(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **10207891 A**

(43) Date of publication of application: **07 . 08 . 98**

(51) Int. Cl **G06F 17/27**
G06F 17/30

(21) Application number: **09006777**

(22) Date of filing: **17 . 01 . 97**

(71) Applicant: **FUJITSU LTD**

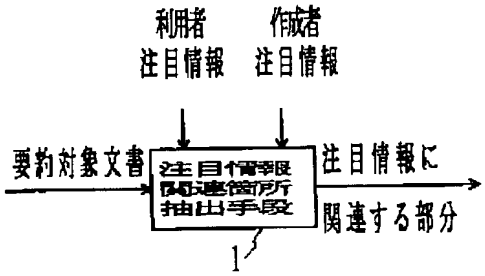
(72) Inventor: **NAKAO YOSHIO**

(54) DOCUMENT SUMMARIZING DEVICE AND ITS METHOD

(57) Abstract:

PROBLEM TO BE SOLVED: To easily grasp how to information noticed by a user is treated in a document only by reading out its summary by extracting parts related to the user's noticed information and writer's noticed information from a document to be summarized based on both the information.

SOLUTION: An extraction means 1 extracts a part related to noticed information as the center part of a summary based on the user's noticed information and writer's noticed information. In this case, an information notice reference for judging that a part including more noticed information is more important is used and the center part of the summary is determined. When two kinds of the noticed information are considered, a summary adopting both of the information requested by the user and the important information in the document, i.e., contents to be written by the writer, is prepared. When either one of the two sorts of the noticed information is weighted, a summary attaching greater importance to the user or the writer is prepared.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平10－207891

(43)公開日　平成10年(1998)８月７日

(51)Int.Cl.[6]	識別記号	ＦＩ		
Ｇ０６Ｆ　17/27		Ｇ０６Ｆ	15/20	５５０Ａ
17/30			15/401	３２０Ａ
			15/403	３３０Ｃ

審査請求　未請求　請求項の数23　ＯＬ　（全 28 頁）

(21)出願番号　特願平9－6777

(22)出願日　平成９年(1997)１月17日

(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号

(72)発明者　仲尾　由雄
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

(74)代理人　弁理士　大菅　義之　（外１名）

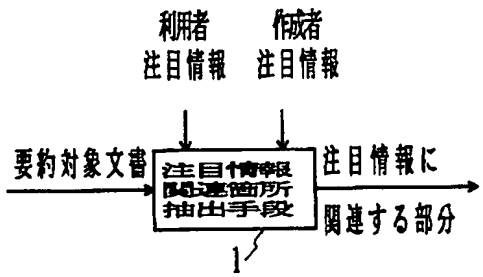
(54)【発明の名称】　文書要約装置およびその方法

(57)【要約】

【課題】　自然言語による、電子化された文書の要約を行う装置に関し、利用者の注目していることと、利用者が既に持っている知識に応じた要約作成を可能とする。

【解決手段】　要約対象文書の中で、要約の利用者が注目している情報としての利用者注目情報と、文書の作成者が注目を促している情報としての作成者注目情報とに基づいて、これら２つの注目情報に関連する部分を抽出する手段１を備え、その抽出結果に応じて要約を作成する。

本発明の第１の原理の説明図



1

【特許請求の範囲】

【請求項１】　計算機可読の文書の選択・閲覧・編集および管理の支援のために文書要約を行う装置において、要約対象文書の内容の中で、要約を利用する利用者が注目している情報としての利用者注目情報、および該要約対象文書の作成者が注目を促している情報としての作成者注目情報に基づいて、該要約対象文書中の該２種類の注目情報に関連する部分を抽出する注目情報関連箇所抽出手段を備えることを特徴とする文書要約装置。

【請求項２】　前記利用者注目情報が、前記要約対象文書の検索のために利用者から入力される質問文の内容であることを特徴とする請求項１記載の文書要約装置。

【請求項３】　前記利用者注目情報および／あるいは作成者注目情報が単語列あるいは重みづけられた単語列の形式であることと、

前記注目情報関連箇所抽出手段が、前記要約対象文書における該単語列内の単語の出現の程度に応じて、前記２種類の注目情報に関連する部分を抽出することを特徴とする請求項１記載の文書要約装置。

【請求項４】　前記文書要約装置において、

前記利用者が興味を有する事柄を利用者嗜好特性としてあらかじめ蓄積する利用者嗜好特性蓄積手段を更に備え、

前記注目情報関連箇所抽出手段が、該利用者嗜好特性蓄積手段の蓄積内容を前記利用者注目情報として利用することを特徴とする請求項１記載の文書要約装置。

【請求項５】　前記文書要約装置において、

前記利用者嗜好特性蓄積手段が複数の利用者のそれぞれに対して利用者嗜好特性を蓄積すると共に、

あらかじめ定められたアクセス制御方式のもとで、前記要約を利用する利用者の利用者注目情報として、異なる利用者の利用者嗜好特性を含む情報を前記注目情報関連箇所抽出手段に与えて前記２種類の注目情報に関連する部分を抽出させる他利用者嗜好特性活用手段を更に備えることを特徴とする請求項４記載の文書要約装置。

【請求項６】　前記作成者注目情報が、通常の流通文書に含まれ、かつ作成者が文書の要点をまとめて提示している情報であって、文書の表題、文書中の章・節および図表の見出し、目次、用語・事項の索引の情報であることを特徴とする請求項１記載の文書要約装置。

【請求項７】　前記文書要約装置において、

複数の要約対象文書に対するそれぞれの作成者注目情報をマージする作成者注目情報マージ手段を更に備え、マージした作成者注目情報に基づいて、前記注目情報関連箇所抽出手段が該複数の要約対象文書中の前記２種類の注目情報に関連する部分を抽出して、

該複数の要約対象文書間の比較情報とすることを特徴とする請求項１記載の文書要約装置。

【請求項８】　前記文書要約装置において、

文書作成時点以後に該文書の作成者あるいは文書管理者によって指定される作成者注目情報を、該作成者注目情報に対応する文書と共に格納する文書格納手段を更に備え、

前記注目情報関連箇所抽出手段が、該文書格納手段に格納されている作成者注目情報を利用することを特徴とする請求項１記載の文書要約装置。

2

【請求項９】　計算機可読の文書の選別・閲覧・編集および管理の支援のために文書要約を行う装置において、利用者がすでに知っている利用者既知情報、および／あるいは該要約作成時点において、過去に利用者に提示された文書に基づいて利用者がすでに知っているとみなせる履歴的既知情報と、該２種類の既知情報以外の情報を区別して使用して要約を作成し、要約の可読性を向上させる要約可読性向上手段を備えることを特徴とする文書要約装置。

【請求項１０】　前記利用者既知情報、および／あるいは履歴的既知情報が既知概念と既知の事柄とによって構成されることと、

前記要約可読性向上手段が、要約内の既知でない概念を減少させ、かつ既知でない事柄については既知性の低いものを優先して要約に取り入れることによって要約の可読性を向上させることを特徴とする請求項９記載の文書要約装置。

【請求項１１】　前記文書要約装置において、

前記概念の既知性が文書内に出現する用語の既知性であることと、

文書内に出てくる用語を認定する用語認定手段と、

該用語認定手段によって認定された用語の既知性を判定する用語既知性判定手段とを更に備えることを特徴とする請求項１０記載の文書要約装置。

【請求項１２】　前記文書要約装置において、

前記事柄の既知性が文書内に出現する用語の組み合わせの既知性であることと、

文書内に出てくる用語の組み合わせを認定する用語組合わせ認定手段と、

該用語組合わせ認定手段によって認定された用語の組み合わせの既知性を判定する用語組合わせ既知性判定手段とを更に備えることを特徴とする請求項１０記載の文書要約装置。

【請求項１３】　前記文書要約装置において、

前記事柄の既知性が文書内に出てくる用語と述語との組み合わせの既知性であることと、

文書内に出てくる用語と述語との組み合わせを認定する用語と述語の組合わせ認定手段と、

該用語と述語の組合わせ認定手段によって認定された用語と述語の組み合わせの既知性を判定する用語と述語の組合わせ既知性判定手段とを更に備えることを特徴とする請求項１０記載の文書要約装置。

【請求項１４】　前記文書要約装置において、

前記利用者が熟知している事柄を利用者知識としてあら

かじめ蓄積する利用者知識蓄積手段を更に備え、
前記要約可読性向上手段が、該利用者知識蓄積手段に蓄積されている利用者知識を前記利用者既知情報として利用することを特徴とする請求項９記載の文書要約装置。

【請求項１５】　前記文書要約装置において、前記利用者知識蓄積手段が複数の利用者のそれぞれに対する利用者知識を蓄積すると共に、
あらかじめ定められたアクセス制御方式のもとで、前記要約を利用する利用者の利用者既知情報として、異なる利用者の利用者知識を含む情報を前記要約可読性向上手段に用いさせる他利用者知識活用手段を更に備えることを特徴とする請求項１４記載の文書要約装置。

【請求項１６】　前記文書要約装置において、
該文書要約装置あるいは該文書要約装置を含むシステムの稼働期間において利用者に提示された文書や要約を利用者の閲覧履歴として保持し、該閲覧履歴を前記履歴的既知情報の基として前記要約可読性向上手段に与える閲覧履歴保持手段と、
該閲覧履歴保持手段に保持されている文書や要約と要約対象文書とを相互参照する文書相互参照手段とを更に備えることを特徴とする請求項９記載の文書要約装置。

【請求項１７】　前記閲覧履歴保持手段が、前記稼働期間を含む長期間に渡る複数の利用者の閲覧履歴をそれぞれの利用者毎に保持することを特徴とする請求１６記載の文書要約装置。

【請求項１８】　前記文書要約装置において、
あらかじめ定められたアクセス制御方式のもとで前記要約を利用する利用者の履歴的既知情報として、異なる利用者の閲覧履歴に基づく履歴的既知情報を含む情報を前記要約可読性向上手段に与える他利用者閲覧履歴活用手段を更に備えることを特徴とする請求項１７記載の文書要約装置。

【請求項１９】　前記文書要約装置において、
要約対象文書の中の各文を、文の述語と該述語に支配される名詞を基本として構成される述語句に分割し、該述語句のうちで他の述語句に依存していない述語句を主述語句とし、該述語句が主題句を含むときは該主題句を分離し、１つの文内または他の文の間での構文的依存構造に従って主題句と主述語句、主述語句と他の述語句との間に依存関係の設定を行い、該設定結果をともなう文書内容を前記要約可読性向上手段に与える文分割・依存関係設定手段を更に備えることを特徴とする請求項９記載の文書要約装置。

【請求項２０】　計算機可読の文書の選別・閲覧・編集および管理の支援のために文書要約を行う装置において、
要約対象文書の内容の中で、要約を利用する利用者が注目している情報としての利用者注目情報、および該要約対象文書の作成者が注目を促している情報としての作成者注目情報に基づいて、該要約対象文書中の該２種類の

注目情報に関連する部分を抽出する注目情報関連箇所抽出手段と、
該抽出結果に対して、利用者がすでに知っている利用者既知情報、および／おるいは該要約作成時点において、過去に利用者に提示された文書に基づいて利用者がすでに知っているとみなせる履歴的既知情報と、該２種類の既知情報以外の情報を区別して使用して要約を作成し、要約の可読性を向上させる要約可読性向上手段とを備えることを特徴とする文書要約装置。

【請求項２１】　計算機可読の文書の選別・閲覧・編集および管理の支援のために文書要約を行う方法において、
要約対象文書の内容の中で、要約を利用する利用者が注目している情報としての利用者注目情報、および該要約対象文書の作成者が注目を促している情報としての作成者注目情報に基づいて、該要約対象文書中の該２種類の注目情報に関連する部分を抽出することを特徴とする文書要約方法。

【請求項２２】　計算機可読の文書の選別・閲覧・編集および管理の支援のために文書要約を行う方法において、
利用者がすでに知っている利用者既知情報、および／あるいは該要約作成時点において、過去に利用者に提示された文書に基づいて利用者がすでに知っているとみなせる履歴的既知情報と、該２種類の既知情報以外の情報を区別して使用して要約を作成し、要約の可読性を向上させることを特徴とする文書要約方法。

【請求項２３】　計算機可読の文書の選択・閲覧・編集および管理の支援のために文書要約を行う方法において、
要約対象文書の内容の中で、要約を利用する利用者が注目している情報としての利用者注目情報、および該要約対象文書の作成者が注目を促している情報としての作成者注目情報に基づいて、該要約対象文書中の該２種類の注目情報に関連する部分を抽出し、該抽出結果に対して、利用者がすでに知っている利用者既知情報、および／あるいは該要約作成時点において、過去に利用者に提示された文書に基づいて利用者がすでに知っているとみなせる履歴的既知情報と、該２種類の既知情報以外の情報を区別して使用して要約を作成し、要約の可読性を向上させることを特徴とする文書要約方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、電子化された自然言語による文書の要約を行う装置に関するものであり、主として、検索された大量の文書の選別や閲覧、あるいは、蓄積された大量の文書の閲覧、再構成（再利用）や管理のプロセスを支援するために、使われることを意図したものである。

【０００２】近年、文書の電子化が進み、大量の文書を

蓄積して再利用することで文書作成のコストを大幅に引き下げることが可能になってきた。また、一方で、技術の高度化に伴うマニュアル類のボリュームの増加と多様化、インターネットに代表される新たな文書流通メディアの出現もあいまって、計算機上で閲覧や再利用のための再構成操作を行える文書の量が爆発的に増加している。

【０００３】このような大量の文書を利用するためには、まず、個々の文書の有用性を素早く判定し、利用目的にあった文書を選別することが重要である。そのためには、文書一覧に文書内容をイメージできるような情報を合わせて提示することが求められる。このような目的にあった情報としては、文書の見出しや抄録などがあるが、見出しが十分に文書内容を表現していない場合や、抄録がない場合も多い。また、特にオンラインで文書を閲覧する場合には、表示可能な文字数が限られるため、例え抄録が作成されていても長過ぎて一覧表示に適さないこともある。そこで、適切な要約を自動的に生成する技術が強く求められることになる。

【０００４】また、文書の再利用を効率的に行うためには、大量の文書を分類・整理して蓄積することが求められる。この場合にも、新たな文書を分類するために当該文書の内容を素早く把握したり、蓄積文書の管理者が分類体系を改良するために分類状況を概観したり、分類体系に通じていない利用者がどのような分類がなされているのかを把握したりすることなどを支援する意味で、やはり適切な要約が望まれる。

【０００５】本発明は、このような用途をもつ文書要約装置において、利用者が何に注目しているのか、また、利用者がどのような知識を既にもっているのかに応じて、要約結果を調整する点に特徴をもつものである。

【０００６】

【従来の技術】従来の文書の要約作成の技術には、大きく分けて２つの方法がある。第１の方法は、文書において重要な部分（通常は、文、段落、節などの文書の論理要素。以下「文」で代表させる。）を認定し、それを抽出することで要約を作成するものである。第２の方法は、要約として抽出すべき情報の型紙を用意して、その型紙の条件にあった文書中の語句を抽出して要約としたり、その型紙によくあてはまる文を抽出して要約とする方法である。第２の方法は本発明との関連性は低いので、ここでは第１の方法を説明する。

【０００７】第１の方法は、さらに、何を手がかりに文の重要性を評価するかによっていくつかの方法に分類される。代表的な方法としては、①文書中に出現する単語の頻度と分布を手がかりとするもの、②文と文とのつながり方や文の出現位置を手がかりとするものの２つがある（その他、文の構文的パターンによって重要性を評価するものなどもあるが、本発明との関連性が低いので省略する）。

【０００８】①の方法、すなわち、文書中に出現する単語の頻度と分布を手がかりとするものは、通常、まず文書中に含まれる単語（語句）の重要度を決定し、次に重要な単語をどれ位含んでいるかで文の重要度を評価し、重要な文を選択するという手順で要約を作成する。単語の重要度を決定する手法としては、ある文書内の出現度数そのままを用いたり、あるいは、一般的な文書集合における単語の出現度数とのずれなどを加味して重みをつけたり、あるいは単語の出現位置に応じて重みをつけたり（見出しに出現する語を重要とみなすなど）する方法が知られている。なお、対象とする単語は、日本語であれば自立語（特に名詞）、英語であれば内容語のみに限るのが通例である。ここで自立語・内容語とは実質的な意味を持つ名詞、形容詞、動詞などの語であり、助詞や前置詞、形式名詞など、専ら構文的役割を示すために使われる語と区別されるものである。なお、日本語の自立語の形式的定義は、独立した文節を構成できる語というもので、前記の説明とは若干のずれがあるが、対象とする単語を自立語に限ることの趣旨は前記の説明中の区別による。

【０００９】このようなものには、例えば以下のようなものがある。特開平６－２５９４２４「文書表示装置及び文書要約装置並びにディジタル複写装置」およびその発明者による次の文献１では、見出しに含まれる単語を多く含む部分を（見出しに関連の深い）重要な部分として抽出することで要約を作成している。

【００１０】文献１：亀田雅之、擬似キーワード相関法による重要キーワードと重要文の抽出．第２回年次大会、pp.97 ～100 、言語処理学会１９９６年３月．

特開平７－３６８９６「文書を要約する方法および装置」では、文書中に現れる表現（単語など）の複雑さ（語の長さなど）から重要な表現の候補（シード）を選び、重要性の高いシードをより多く含む文を抽出することで要約を作成している。

【００１１】特開平成８－２９７６７７「主題の要約を生成する自動的な方法」では、文書内の単語の出現頻度が大きい順に「主題の用語」を認定し、重要な「主題の用語」をより多く含む文を抽出することで要約を作成している。

【００１２】特開平成６－２１５０４９「文書要約装置」では、検索結果と質問文との関連性判定などによく用いられるベクトル空間モデルを適用して、文書全体の特徴ベクトルと最も類似した特徴ベクトルを持つ文や段落から文を選択していくことで要約を作成している。ここで、ベクトル空間モデルとは、キーとなる単語ごとに、あるいは単語の意味成分ごとに、次元（軸）をわりあて、文書や質問文におけるそれらの単語の出現の有無あるいは出現頻度の値の列（特徴ベクトル）で、文書や質問文の特徴を表現するものである。

【００１３】②の方法、すなわち、文と文とのつながり

方や文の出現位置を手がかりとするものとは、順接・逆接、展開などの文の接続関係（結束関係と呼ぶ場合もある）や、文が出現している文書中の位置などをもとに、文の（相対的）重要性を判定し、重要な文を選択するものである。例えば、特開平７－１８２３７３「文書情報検索装置及び文書検索結果表示方法」およびその発明者らによる次の文献２、別の著者による文献３がある。

【００１４】文献２：住田一男、知野哲朗、小野頭司、三池誠司、文書構造解析に基づく自動抄録生成と検索提示機能としての評価・電子情報通信学会論文誌、Vol.J78-D-II,No.3,pp.511～519,１９９５年３月.

文献３：山本和英、増山繁、内藤昭三、文書内構造を複合的に利用した論説文要約システムＧＲＥＥＮ．情処研報ＮＬ－９９－３，情報処理学会１９９４年１月.

以上のように文書全体の要約を作成する技術の他に、個々の文書の有用性の判定を支援するための技術として、利用者の注目している部分を提示する技術もある。周知の技術として、ＫＷＩＣ（Keyword In Context）と呼ばれる検索語の周囲を表示する方法や、それに類似した検索語の近傍表示の方法が広く使われている。

【００１５】また、論文において研究の背景となる事情を述べた部分だけを提示したり、新聞の第一パラグラフだけを提示するなど、利用者の目的に応じて特定の部分だけを提示する方法もある。例えば、前掲の特開平成７－１８２３７３や文献３、別の著者による次の文献４，５がある。ただし、これらは、分野に特有な文書構成の類型や言い回しなどを手がかりとして、文書の論理構造上で特別な機能を持つ部分を選択するものであり、利用者が注目している内容に着目して、その内容と関連性の高い部分を提示しているわけではない。

【００１６】文献４：神門典子．複数領域における日本語原著論文の機能構造分析：構成要素カテゴリの自動付与．Library and Information Science，No.31，pp.25～38，１９９３年.

文献５：神門典子、原著論文の機能構造の分析とその応用．図書館学会年報，Vol.40,No.2,pp.49 ～61，１９９４年６月.

要約の可読性を低下させる要因としては、冗長な表現、利用者の知らない用語（未知の用語）の出現、解消されていない照応表現（anaphoric expression：日本語ならいわゆるコソアド語など）の出現などがある。

【００１７】このうち、冗長な表現については、修飾要素と被修飾要素の語彙的な性質や関係、修飾要素と被修飾要素の距離などによるヒューリスティックにより、余分な修飾要素を削除する手法が知られている。例えば、前掲の文献３では、日本語の新聞記事中の文において同じ名詞に対して２つ以上の修飾要素があった場合、前の修飾要素を削除するというヒューリスティックが示され、また、同じ著者らによる次の文献６では、同一事件に関する一連の記事で、特有の言い回しから導入部と認定された部分で出現する名詞の７割以上が以前の記事で既出の場合にその導入部を削除するというヒューリスティックが示されている。

【００１８】文献６：船坂貴浩，山本和英，増山繁．冗長度削減による関連記事の要約．情処研報ＮＬ－１１４－７，情報処理学会，１９９６年７月.

未知の用語の出現については、用語の定義や説明をしている部分が文書中にあれば、それを要約に含めればよいことは自明である。このような部分を探すには、初出の部分あるいは用語の定義を示すマーク（日本語なら「とは」など）の付いた当該用語が出現する箇所を探せばよいことも、学校教育レベルの知識である。

【００１９】照応表現についても、その先行詞（antecedent）を探し、照応表現を先行詞で置き換えたり、先行詞を含む部分を要約に含めれば、要約が理解しやすいものとなることは自明である。照応表現の先行詞の同定に関しては、センタリング（centering ）と呼ばれる手法などが知られている。これは、後続の文で照応表現になりやすい要素（センタ）を構文的役割などに応じて優先度つきで認定しておき、後続の照応表現の現れ方による制約を加味して、先行詞をセンタの中から選択するというものである。なお、類似の手法で、センタと同様な概念を焦点（focus ）と呼ぶものもある。ただし、いずれにしても完全な結果を得ることができる技術ではない。センタリングの手法については次の文献７，８がある。

【００２０】文献７：Megumi Kameyama. A property-sharing constraint in centering. In Proceedings of the 24th Annual Meeting of Association for Computational Linguistics, pp. 200～206,１９８６年.

文献８：Susan E.Brennan, Marilyn W. Friedman, and Carl J. Pollard. A centering approach to pronouns. In Proceedings of the 25th Annual Meetingof Association for Computational Linguistics, pp. 155～162,１９８７年.

前傾の特開平７－１８２３７３や、同一発明者らによる特開平７－４４５６６「抄録作成装置」では、このような手法を実装して用語の定義箇所や照応表現の先行詞を推定し、元の用語や照応表現からハイパーテクスチュアル（hyper-texutual）なリンクを設定するなどして利用者の便を図っている。

【００２１】

【発明が解決しようとする課題】大量の文書から有用な文書を選別するためには、利用者にとっての文書の有用性を素早く判定できるよう、利用者の求める情報を文書の作成者がどのように取り扱っているのかを示すことが重要である。検索システムでは、利用者が得たいと思っている情報は、質問文やキーワードによる検索式として表現されることが多い。しかしながら、質問文や検索式中の単語が検索対象の文書の中に見つかったからといって、そこに利用者の求める情報が書かれているとは限ら

ない。たとえば、特許公報を「翻訳」というキーワードで検索すると、利用者の得たい情報は「自然言語の文の翻訳」に関する特許情報なのに、検索結果には「機械語の翻訳」に関する情報が大量の特許が含まれることがある。この場合、「翻訳」という語がどういう文脈で使われているのかを提示すれば、ある程度、文書の選別を支援することが可能である。このような目的で、前節に挙げたＫＷＩＣを用いることもできるが、物理的な近傍を表示するだけなので論旨の流れなどを把握することが難しく、簡潔で要を得た要約を提示することは必ずしもできない。

【００２２】この観点からみると、前述のように、従来の要約作成の技術では専ら文書中での文の重要度のみを要約中に文を含めるかどうかの基準としており、利用者の注意がいずれに向けられているのかが考慮されていない、という問題がある。そのため、例えば検索システムに入力したキーワードが、言語学の文献の用例の部分で照合されて取り出されてしまった場合などでは、自動作成された要約中にはキーワードが含まれず、利用者の入力とどのような関連がある文献なのかを判定しづらくなる場合などが考えられる。

【００２３】言語学の文献に関連する問題点についてさらに説明する。言語学の文献は、言語の形式的な性質を議論するもので、そこで取り上げられている用例の内容については言語学と関係している必要はない。例えば、「象は鼻が長い」や「僕はうなぎだ（料理の注文の意）」というのは国語学では有名な例文である。動物のことを調べようとしている時に、このような例文で出ている言語学の文献が検索されてしまう可能性がある。言語学の文献のために、文献中に含まれる語彙の頻度分布などをとれば、動物に関する語彙は少なく、例えば「象」はあまり重要でない語彙であると計算されてしまう。そうすると、検索結果の表示に語彙の頻度分布などに基づいて自動生成した要約を用いる場合、このような例文は要約中に含まれにくくなり、不都合が生じる可能性がある。つまり、「象」というキーワードを入力した時に、このような言語学の文献が検索され、それにも関わらず、検索結果の表示（自動生成した要約）の中には「象」が含まれず、何でこのようなものが検索されたのかが理解できなくなる可能性があるということになる。逆に、キーワードの近傍表示だけしかしない場合には、用例の部分だけが表示され、どういう趣旨の文献なのかが理解できなくなることもある。

【００２４】もう一つの問題点として、従来の要約作成の技術には、利用者の知識レベルに合わせて要約を作成する手段を備えていない、という問題点もある。利用者の知識レベルは、利用者ごとに大きく異なる可能性があり、特に専門的な用語に関する知識がどの程度あるかに合わせて要約に用語に関する定義や説明の記述部分を含めるかどうかを切り換えないと、知識レベルの高い利用者にとっては冗長な要約になったり、知識レベルの低い利用者にとっては理解し難い要約になってしまったりすることになる。

【００２５】以上のことから、本発明は要約作成において、利用者の注意の方向が考慮されていない、利用者の知識レベルが考慮されていない、という２つの問題点を解決し、それらを統一的に扱う手段を提供することを目的とする。

【００２６】

【課題を解決するための手段】図１、および図２は本発明の原理説明図である。これらの図は、自然言語による電子化された文書の選別・閲覧・編集、および管理の支援のために、文書要約を行う文書要約装置の原理を説明するものである。

【００２７】図１は本発明の第１の原理の説明図である。同図において注目情報関連箇所抽出手段１は、利用者の注目情報と作成者注目情報とに基づいて、要約対象文書中でこれら２つの情報に関連する部分を抽出するものである。ここで利用者注目情報とは、要約対象文書の内容の中でその要約を利用する利用者が注目している情報であり、また作成者注目情報とは要約対象文書の作成者が利用者に対して注目を促している情報である。

【００２８】図２は本発明の第２の原理の説明図である。同図において要約可読性向上手段２は、利用者既知情報と履歴的既知情報とに基づいて、これら二種類の既知情報以外の情報と二種の既知情報とを区別して用いて要約を作成し、要約の可読性を向上させるものである。ここで利用者既知情報とは、要約対象文書の内容の中で、その要約を利用する利用者がすでに知っている情報であり、また履歴的既知情報とは、要約作成時点において、それ以前に利用者に提示された文書に基づいて利用者がすでに知っていると見なせる情報である。

【００２９】後述する本発明の実施例においては、図１と図２とによって説明される２つの原理を同時に用いて、要約の作成が行われる。まず図１においては、利用者注目情報および作成者注目情報という２種類の注目情報に基づいて、注目情報に関連する部分が要約の中心部分として抽出されることになる。ここではこれらの注目情報を多く含む部分ほど重要であると判定する情報の注目性基準が用いられて、要約の中心部分が決定される。

【００３０】２種類の注目情報を考慮することによって、利用者が求める情報と、文書において重要な情報、すなわち作成者が書こうとしていた内容の双方を取り込んだ要約を作成することができる。これら２種類の注目情報のいずれかに重みをつけることによって、利用者注目情報だけを重視した要約から、文書における重要性だけを重視した要約まで、目的に応じた要約を作成することができる。あるいはこれら２種類の注目情報を同等に扱うことによって、利用者が求める情報と、作成者が書こうとしていた内容との双方をバランスよく抽出した要

約を作成することもできる。

【００３１】図２においては、利用者既知情報と履歴的既知情報との２つに基づいて要約が作成される。これは利用者の知識レベルに合わせて要約を作成することを意味し、これによって要約の可読性を向上させることができる。この可読性の向上のために情報の既知性基準が用いられる。情報の既知性基準は、例えば概念の既知性基準と事柄の既知性基準との２つの基準を意味する。

【００３２】概念の既知性基準とは、要約を構成する要素概念、特に主題に関する要素概念が原則として既知でなくてはならないという基準である。ここで要素概念とは、要約に含まれる個々の語句が表す概念のことである。言い換えれば、要約に出力する用語（主として名詞）は、原則として利用者にとって既知でなくてはならないという基準である。本発明においては、この基準に基づいて、利用者に理解できない用語に関しては、例えば必要な説明を追加して要約が作成される。

【００３３】事柄の既知性基準とは、例えば文書の中に出現する用語の組み合わせに関するものであり、その組み合わせ全体で述べられている事柄（事実、あるいは命題）については、既知性の低いものほど優先されて要約に取り入れられる。

【００３４】この事柄の既知性基準によって通常の場合、すなわち独立した文書１つを要約する場合などは、要約の中で同じ事柄が何度も出力されるのが抑制されることになる。また関連する複数の文書を一括して要約する場合、特に同一事件に関する一連の記事や、記載コラムなどをまとめて、要約対象の文書の間の関係を明らかにして提示できる場合には　同一の事柄に関する記述を削減するための基準として用いられる。

【００３５】この事柄の既知性基準は、前述の情報の注目性基準によって注目情報の含み方が同程度と判定された文が複数あった場合には、事柄の既知性の低い方を選んでその低い方の文の内容が要約に含められるという意味で、弱めの制約となるものである。すなわち、概念の既知性基準が「（原則として）既知でなければならない」のに対して、事柄の既知性基準は「既知性の低いものが優先される」だけであるために弱めの制約となる。

【００３６】このように本発明においては、情報の注目性基準と情報の既知性基準という２つの基準を用いて、要約の作成が行われる。

【００３７】

【発明の実施の形態】図３は本発明の文書要約装置の構成を示すブロック図である。同図において、文書要約装置は要約プロセス制御部１０、文書構造解析部１１、文解析部１２、文分割・依存関係設定部１３、文選択部１４、要約整形部１５を基本構成要素として備えている。このうち本発明にとって特徴的な構成要素は、文分割・依存関係設定部１３と、文選択部１４である。

【００３８】要約プロセス制御部１０は、利用者と装置とのインタフェースとなると共に、文書要約装置の動作全体を制御するものである。利用者との間のインタフェースとしては、利用者が注目している情報、すなわち利用者注目情報や、要約作成に関する要望などの入力を受け取り、文書要約のプロセスを適切に起動し、要約結果を利用者に出力することになる。利用者注目情報の代表的な形式は利用者から入力される質問文であるが、求める情報に関するキーワードや、読書案内に載っている紹介文の形式とすることも可能である。

【００３９】要約作成に関する要望、すなわち要約作成に関する制約情報としては、利用者から必須出力要素が指定されると共に、その他の制御命令が与えられる。ここで必須出力要素とは、例えば見出しのように、要約の中に必ず含めるべき要素である。その他の制御命令としては、注目情報や既知情報としてどのような情報を利用するか、またそれらの情報をどのように利用するか、望ましい要約の長さ、要約処理でどのような単位を基本として要約を作成するかなどがある。この基本単位としては、通常は文、あるいは述語句が用いらる。

【００４０】また、要約プロセス制御部１０では、複数文書の比較を支援するために本装置を用いる場合に、まず個々の比較対象文書に関する文書構造認識解析および文解析までの処理を行い、その出力を集計して作成者注目情報のマージし、マージした注目情報にもとづいてそれ以降の要約処理（文分解・依存関係設定処理、文選択処理、要約成形処理）を行うよう、各処理部の動作の制御および処理経過の記憶も行う。

【００４１】図３において、文書要約装置のメモリには、利用者の嗜好特性１６、利用者の知識１７、および閲覧履歴１８が蓄積されると共に、例えば他のメモリに入力文書（群）１９が格納される。

【００４２】利用者の嗜好特性１６は、利用者が興味を持っている事柄を蓄積するものである。ここには利用者が、例えば自己紹介の際に使うような趣味の説明文や、利用者が興味を持った文書そのままを蓄積したり、そのような文書の中から出現頻度の大きいキーワードを抽出して保存したり、利用者が検索に際してよく使うキーワードや質問文を保存しておいてもよい。

【００４３】利用者の知識１７は、利用者がよく知っている情報を、利用者既知情報として蓄積するものである。ここには例えば利用者が知っている専門用語のリストなどが蓄積される。

【００４４】閲覧履歴１８は、利用者がどのような文書や要約をいつ頃閲覧したかという履歴を蓄積するものである。入力文書（群）１９、基本的には要約対象文書を格納するものであり、通常はどのような形式の電子化文書でも用いることができる。具体例としては、電子出版等で用いられている文書構造記述言語のＳＧＭＬ（スタンダードジェネラライズドマークアップランゲージ。ＩＳＯ８８７９）を用いればよい。要約対象文書に対し

て、例えば作成者あるいは文書の管理者によって、文書の作成時点以後に指定された作成者注目情報を、文書と対応させて蓄積することもできる。

【００４５】メモリの内容としての利用者の嗜好特性１６、すなわち利用者注目情報、利用者の知識１７、すなわち利用者既知情報、閲覧履歴１８、すなわち履歴的既知情報、入力文書（群）１９すなわち要約対象文書は要約プロセス制御部１０によって管理され、要約の作成に使用される。

【００４６】文書構造解析部１１は、要約プロセス制御部１０から要約対象文書や必須出力要素の指定内容などを受け取り、文書の構造を解析し、文書内容を文解析部１２に出力すると共に、依存関係付文書構造情報を文分割・依存関係設定部１３に与えるものである。

【００４７】文書構造解析部１１は、まず文書の書式やマークマップ情報などから、見出しや本文というような文書の論理的構成要素を認定し、例えば見出しと本文、あるいはそれと同様の関係を持つ要素として列挙構造の項目名とその内容などを対応づけて、構成要素間の依存関係を抽出する。この依存関係では、例えば本文内の要素が従属ブロック、見出し内の要素が依存先とする。

【００４８】文書構造解析部１１は、作成者注目情報として文書の論理的構成要素、すなわち章、節・図表などの見出しや、目次、用語や事項の索引などを用いる場合に、作成者注目情報にあたる部分を認定し、作成者注目の印をつけて、文解析部１２および文分割・依存関係設定部１３を介して文選択部１４に与える。要約プロセス制御部１０によって必須出力要素が指定された場合には、該当する部分に必須出力の印をつけて、同様に文選択部１４に与える。文解析部１２が１つの文を単位としてしか処理できない場合には、文の認定も行うことになる。

【００４９】文書構造解析部１１による具体的な処理は、文書の種類、例えば単なる自然言語の文書か、構造化された文書（例えばＳＧＭＬ文書）かなどによって異なり、本発明にとって本質的なものではないので、その詳細の説明は省略する。

【００５０】文解析部１２は、文書の内容を文書構造解析部１１から受け取り、それに含まれる単語を認定し、単語の出現位置や品詞情報をつけた単語列の形で文書内容を文分割・依存構造解析部１３に出力するものである。また、利用者注目情報が質問文などの自然言語の形で与えられた場合には、自然言語の利用者情報からも同様に単語列を作成し、文選択部１４に出力する。具体的な処理の方法としては、形態素解析法として各種のものが知られており、それを用いればよいので説明は省略する。なお、単語列に付与される出現位置とは、文構造解析部１１から出力される依存関係付文書構造と単語列とを対応付けるものであり、文分割・依存構造解析部１３では文書構造解析部１１で設定された文書の構成要素間の依存関係を述語句間の関係に変換するために使われ、文選択部１４では文構造解析部１１で設定された必須出力の印および作成者注目の印に従い、必須に出力する述語句や作成者注目情報に対応する注目概念を認識するために使用される。

【００５１】文分割・依存関係設定部１３は、文書構造解析部１１から出力される依存関係付文書構造情報と、文解析部１２から出力される出現位置や品詞情報がつけられた単語列、および要約プロセス制御部１０から出力される既知概念を用いて、後述する文分割処理と依存関係設定処理を行い、文選択部１４に対して依存関係付述語句列（述語句リスト）を出力するものである。

【００５２】文選択部１４は、要約プロセス制御部１０や文解析部１２から出力される注目情報と、要約プロセス制御部１０から出力される既知の事柄を示す情報に従って、文分割・依存関係設定部１３から出力される依存関係付述語句列に対して後述する文選択処理を実行し、要約に含めるべき重要な述語句（文）を選択して、後述する選択結果リストを作成するものである。

【００５３】ここで、質問文などの自然言語の注目情報が文解析部１２に与えられ、その他の注目情報が文解析部１２を経由することなく、直接文選択部１４に与えられる理由を説明する。本実施例においては、後述するように文の重要度としての注目情報量は、単語、例えば名詞を単位に計算される。そこで自然言語文として注目情報が与えられた場合には、その自然言語文を単語に分割する必要がある。質問文や文書から取り出された見出しなどが、文解析部１２を経由して文選択部１４に渡されるのはそのためである。

【００５４】一方、例えば利用者の嗜好特性１６として蓄積されている利用者注目情報などは、あらかじめ文解析を行った後に適切な形式でメモリに格納することが可能であり、この場合に文解析部１２を経由することなく、文選択部１４に直接に与えることができる。なお本実施例では自然言語と無関係な情報は利用しないが、メモリへの格納形式は後述する意味ネットワーク表現であったり、フレーム表現であったりしてもよく、蓄積された情報は単なる自然言語に限られない。

【００５５】文選択部１４で用いられる注目情報としては、代表的には名詞のリストが与えられる。注目度の高い名詞には注目度に対応する重みを与えることもできる。また名詞以外の自立語（動詞や形容詞など）を注目情報として与えたり、名詞と用言の組として与えることも可能である。以下の説明では、重みなしの名詞リストが注目情報として与えられる場合を中心に実施例を説明する。なお述語句の選択においては、文分割・依存関係設定部１３の処理によって設定された概念の既知性基準に適合した述語句（文）間の依存関係に違反しないように選択処理が行われるため、概念の既知性基準にも適合した選択結果リストが作成されることになる。

15

【００５６】要約整形部１５は、文選択部１４によって選択された文を元の文書における出現順に並べ、必要に応じて抽出されなかった文の存在を表す印や、段落の境界などを挿入し、要約を読みやすい形式に整形する。履歴的既知情報への依存関係が設定されている場合には、ハイパーテクスチュアルな関係を設定することもできる。

【００５７】ここで既知概念と既知の事柄について更に説明する。既知概念は基本的には実質的な意味を持つ単語としての内容語のリストである。例えば富士通が何をしている会社なのか知っている場合には、既知概念として「富士通」を与える。そうすると要約対象文書が富士通は日本の計算機メーカである。その富士通で今・・・が行われようとしている。・・・のように始まっている時、通常は第２文の先頭の「その富士通」に「その」という照応表現が含まれているため、第１文も要約に含むように処理される。しかしながら「富士通」は既知の概念であることが知らされ、また第１文は後述するような名詞文であり、富士通の紹介（属性の定義）をしているだけであると計算機の処理でも簡単に判断可能なので、本実施例では第１文は抽出されないことになる（第２文の「その」は削除される。）。

【００５８】但し、このような単純な方法をとる場合には、次のような文書を対象にするときに不都合がおきることがある。「富士通はもともと交換機をつくる会社だった、その富士通が、大型計算機で世界二位の地位をしめるようになり、今ではパソコンメーカとして知っている人の方が多いくらいだろう。だからＮＴＴと富士通の組み合わせを不思議に思う人もいるかもしれないが、ＮＴＴと富士通の関係は浅からぬものなのである。・・・」例えば利用者がパソコンの富士通しか知らない場合、第３文を要約に含める時には、第１文も共に含めないと理解が困難となる。これを回避するためには、利用者が「富士通」についてどのような事柄を知っているのかまで指定する必要がある。例えば「富士通は日本の計算機メーカである」、「富士通はパソコンを作っている」ことまでは知っているというように指定しなければならない。更に一歩進むと、「富士通は交換機を作っていた」ことは知っていても、「富士通が今でも交換機を作っている」のか、あるいは「富士通がこれからも交換機を作り続ける」のかは知らないということがあり得る。これが本実施例において事柄の既知性を取り扱うことの意味である。

【００５９】このような意味で、本実施例では既知概念としては主に専門用語のリストを与える。専門用語は専門的な概念に名前をつけたものであり、分野を誤らない限りは、それが既知かどうかは容易に決められる。既知の事柄に関する知識としては、上の例のような単文の形、または単文の内容に相当するものを、フレーム表現など各種の形式で表現したものを与える。

16

【００６０】また既知の概念は、補足的・説明的な記述が必要であるか否かの判定に用いられるために、既知概念は文分割・依存関係設定部１３に与えられる。すなわちある部分を要約に含める時、それを説明しているような別の部分も含めるか否かの判定が文分割・依存関係設定部１３の役割であって、ある単語が既知の概念であるかどうかは、その判定に強く影響を与えるためである。

【００６１】文書要約装置の構成の説明に続いて、この装置内での本発明に特有の構成要素としての文分割・依存関係設定部１３と、文選択部１４の処理について説明する。図４は、文分割・依存関係設定部１３による、文分割・依存関係設定処理の詳細フローチャートである。

【００６２】図４においては、文分割処理と依存関係設定処理とが行われるが、文分割処理は最終的に依存関係付述語句リストを作成する文分割・依存関係設定部１３の処理の前半部分である。但し図４においては、文を認定し文と文の間に依存関係を設定する処理の途中に文の内部構造を解析し、文内部の述語句間に依存関係を設定する処理が挟み込まれているので、文分割処理と依存関係設定処理とを単純に２分することはできない。

【００６３】文分割処理とは、文解析部１２によって単語列に変換された文書の内容を、文選択部１４による文選択処理における基本単位（述語句あるいは文）に分割して、分割された基本単位を要素とするリストとしての述語句リストを作成する処理である。この文分割処理は、図４において、ステップＳ２による先頭の文を取り出すという処理と、点線で囲まれた述語句への分解処理の中で実行される。

【００６４】図４において処理が開始されると、まずステップＳ１で最終的に作成されるべき述語句リストの内容がクリアされ、ステップＳ２で先頭の文が取り出され、ステップＳ３で文が取り出せたか否かが判定される。文書からの文の取り出しは、例えば見出しをそれだけで１つの文と見なすなど、文書の論理構造も考慮しながら、ピリオドなどの文の終了マークを手がかりとして行うことができる。

【００６５】文が取り出せたと判定されると、ステップＳ４で文の構文的依存構造の解析が行われる。この構文的依存構造を求める方法としては、句構造文法によるもの、係り受け解析によるものなど様々な方法が知られているので、それらのいずれかを用いればよい。

【００６６】構文的依存構造の解析結果を基にして、ステップＳ５で取り出された文の述語句（単文）への分解が行われる。述語句とは、１つの述語とそれに支配される名詞（主語を含む）を基本として構成される句であり、文に含まれる単文に相当する。日本語なら用言、英語なら動詞などのような依存構造中の述語が取り出され、それに依存している要素のうち述語でないものを加えたものが述語句である。接続詞や助詞、前置詞などの機能語や、機能語に相当する表現はその前後の自立語

（内容語）とまとめておけばよい。

【００６７】なお単独で名詞を修飾する形容詞などの修飾要素は、被修飾要素と一緒にまとめてもよく、独立した述語句としてもよい。但し好ましくは修飾要素の語彙的性質や、修飾要素と修飾要素の組み合わせの種類などによって、独立した述語句とするかどうかを決定する。

【００６８】依存構造の解析方法としてどのような方法を用いるとしても、述語句への分解のコストはかなり高くなるが、述語句を単位として要約することにより、長い文の場合などでも簡潔な要約が生成可能となり、また高度な意味処理を行うために既知の事柄を図５に示すような格フレーム形式（フレームとは、属性名（スロット名）と属性値（フィラー）の組の列であり、知識表現法として周知のものである。）で与えたり、図６に示すような意味ネットワーク表現で与える場合などにおいて、既知の事柄情報と要約の単位との照合が簡単になるというメリットがある。

【００６９】なお図６の意味ネットワークにおいて、アンダーラインは意味を表す単位となるシンボルを示し、アンダーラインのない矢印付の単語は関係を表す。図６ではシンボルが日本語で表現されているが、例えば“発表する”に対して英語の“announce”を対応するシンボルとしてあらかじめ定義しておくことにより、日本語だけでなく英語に対する情報としても使用することができる。

【００７０】以上の理由から、要約の単位としては必要に応じて述語句、あるいは文のいずれかを使い分けることが望ましい。文を単位として要約を行う場合には、図４において点線で囲まれた述語句への分解処理、すなわちステップＳ４～Ｓ１２を省略することができる。この部分には、前述の述語句への分解の部分と、次に述べる述語句間の依存関係の設定処理が含まれている。

【００７１】次に図４における依存関係設定処理について説明する。依存関係設定処理は、文分割・依存関係設定部１３における処理、すなわち図４の処理の後半部分である。この処理では、図３において要約プロセス制御部１０から与えれらる既知概念のリスト、および文書構造解析部１１からの出力であり、文書構造の上から推定される構成要素間の依存関係に基づいて、概念の既知性基準による制約を、述語句リスト中の要素（述語句あるいは文）の間の依存関係の形で付与する処理が実行される。この処理は、文分割処理によって取り出された文毎に順次実行される。

【００７２】構成要素間の依存関係は、他の文や句（依存先の文や句）を一緒に抽出した方が要約が読みやすくなるような文や句（従属文・句）に対して設定されるものである。そのような依存関係の設定対象には以下のようなものがある。

（１）文に含まれる従属文

注目要素を含まない従属文を省略し、注目要素を含む従属文を主文と一緒に要約に含めるようにする場合に、従属文から主文への依存関係を設定する。長い文の多い特許公報などを要約する場合に有効である。

（２）前後の文に強く依存している文

逆接の接続詞（「しかし」など）を文頭に含む文などについて、注目要素を含まない場合には省略し、注目要素を含む場合には前の文も必ず一緒に要約に含めるようにする場合に、従属文（この例では後ろの文）から依存先の文（この例では前の文）への依存関係を設定する。短い文を積み重ねて書かれた論文などの場合に有効である。

（３）見出しのついている部分に含まれる文

章節などに分かれている文書で、章節内の文に注目要素が含まれている時に章節の見出しを必ず一緒に要約に含めるようにする場合に、章節内の文から章節の見出しへの依存関係を設定する。マニュアルなどの構造化された長めの文書から、知りたい事柄がどこに書いてあるのかを調べる場合などに有効である。

（４）主題となっている語句が既知でない文

特に動詞（の過去形「～た」）の文の主題となっている語（文頭の「～は」など）が要約中で初出の場合、その語が初めて出現した文（初出の文）も一緒に要約に含めるようにする場合に、主題が既知でない語の文から初出の文への依存関係を設定する。経済関係の雑誌などの長めの記事の場合に有効である。これは、概念の既知性基準による処理の一つである。

【００７３】要約対象がかなり長い場合には、初出の文のかわりに、既知でない主題の語が、作成者注目情報とともに現れている文で近傍に出現するものを用いたり、既知でない主題の語が、必須格（「が」「を」「に」）を伴って出現している文で近傍に出現するものを用いた方がよい場合もある。なお、語の一致の判定は、同一表記の語を一致とするだけでなく、略語と正式名称との一致や同（類）義語動詞の一致などについても行うことが望ましい。英語でも、固有名詞が主語となっている動詞文などについて、同様な処置が有効なことがある。

（５）照応表現を含んでいる文

日本語のコソアド語（「これ」「この」「こう」など）や英語の３人称代名詞などの照応表現が登場する文に注目情報が含まれていれば、先行表現が含まれている部分も一緒に要約に含めるようにする場合に、照応表現を含んでいる文から先行表現を含んでいる部分への依存関係を設定する。照応表現を含んでいる文にペナルティを与え、選択されにくくすることの方が有効な場合もある（特に、日本語の名詞文（「～は～だ」）や形容詞文（「～は～い」）など）。英語の場合や翻訳調の日本語文書の場合に有効である。ここで、照応表現を含んでいる文にペナルティを与えるということについて、名詞文・形容詞文に関する例をあげて補足説明する。「名詞文」「形容詞文」とは、形式的には文の述語が「名詞＋

だ／です」、あるいは形容詞・形容動詞の文のことで、典型的には主題を表す「～は」の文節（俗にいう主語）を含むものである。内容的には、主題となっているものの性状や、主題に関する話者の判断を表している文であり、品定め文などとも呼ばれる（例えば三上章『現代語法新説』くろしお出版１９７２年）。本明細書で「名詞文」「形容詞文」というのは、内容的な観点からの表現で、必ずしも「名詞＋だ」や形容詞を述語としている文全てを対象としているのではない。逆に述語が動詞であっても「彼はよく働く」のように性状を表現する文であれば同様に考えることができる。

【００７４】名詞文・形容詞文は、文書の中では図７に示す例（翻訳記事）のように、すぐ後で述べることを導入するために用いたり、話題のつなぎを示すために使われることがよくある。このようなものは表現が抽象的であったりして、前後を見ないと具体的にどういう事柄を述べようとしているのかを掴めないものが多く見られる。そのため、そういう文だけを抜粋しても、表現の意図の理解が難しく、要約に採りあげるには意味のないことがある。例えば、図７の初めの例（アンダーライン部）であれば（典型的な形容詞とはいえないが）、後続の文を採りあげた方が、その記事が何について書いたものなのかを把握する助けとなる。次の例では、その直前の文およびそのさらに前の文を採りあげた方がよい。

【００７５】本発明のねらいは、新聞や雑誌の記事やマニュアル類から必要な事柄（知識）を素早く見い出すことにある。そこで、図７の例のような、話題にまとまりをつけたり、展開のつなぎにするような文の価値は低くなる。特に照応表現がこのような文に現れた場合には、前後で述べている事柄にニュアンスを追加するようなものが多いので、ペナルティを与えて選択されにくくする、というのが照応表現にペナルティを与える趣旨である。

【００７６】以上の（１）～（５）のような文に対応して、依存関係またはペナルティを設定することによって、要約の可読性を高めることができるが、実際にはそれ相応の計算コストが必要である。特に（４）の主題となっている語句が既知でない文、（５）の照応表現を含んでいる文の依存関係に対しては、その関係を完全に処理する技術が存在せず、不適当な依存関係が設定されるとかえって可読性を損なう場合が考えられる。そこで本実施例においては、基本的な処理の流れを示すために、（１）の従属文、（３）の見出しのついている部分に含まれる文、（４）の主題となっている語句が既知でない文について依存関係を設定し、（５）の照応表現を含んでいる文についてはペナルティを設定する場合について処理を説明する。なお、ペナルティについては、後述する文選択部１４の処理において、文選択の基準となる情報量を、照応関係を含む文については通常の場合より減少させることによって、照応関係を含んでいる文が選択されにくくするような処理が実行される。

【００７７】図４のステップＳ５において文が述語句（単文）に分解された後に、ステップＳ６で構文的依存構造において依存関係にある述語の間に依存関係が設定され、ステップＳ７で他の述語句に依存していない述語句が主述語句として設定され、ステップＳ８で主述語句が述語句リストに追加される。なお前述のように文単位に処理を実行する場合には、これらの処理は省略され、単に文全体が主述語句とされる。主述語句とは、後続の処理において文と文の間の依存関係を設定する際に、依存先となるものである。

【００７８】図８は述語句への分割と依存関係設定の例である。同図(a)において、文１における述語句２は述語句１に依存しているという依存関係が設定される。また文２に対しても、同様に述語句２が述語句１に依存するという依存関係が設定される。いずれの文においても述語句１が主述語句である。このように、依存関係は、構文的依存構造における述語間の関係が直接的な場合（文１）あるいは間接的な場合（文２）のいずれについても同様に設定される。すなわち、文１においては述語句２の述語「引いた」は接続助詞「ので」を介して直接的に述語句１の述語「休んだ」と関係しており、一方文２においては述語句２の述語「送ってくれた」は名詞「手紙」を介して間接的に述語句１の述語「しまった」と関係しているが、どちらも同じように依存関係が設定される。

【００７９】図４のステップＳ８で主述語句が述語句リストに追加された後に、ステップＳ９～Ｓ１２で文の代表句が決定される。代表句とは、その文に依存先がある場合に後述のステップＳ１４で設定される依存関係の起点となる句のことである。文を単位として処理を行う場合には、ステップＳ９～Ｓ１２の処理も不要であり、文全体を代表句（かつ主述語句）とすればよい。

【００８０】まずステップＳ９で文に主題句が存在するか否かが判定される。主題句がある場合には、ステップＳ１０で主題句が分離され、主題句と主述語句との間に依存関係が設定される。主題句とは、日本語の主題マーカ（「は」など）のついた体言句のことである。

【００８１】図８(b)は主題句分離後の依存関係を示している。文１に対しては、主題句は「太郎は」であり、述語句２は述語句１に依存し、述語句１は主題句に依存するという関係が設定される。文２に対しては、主題句は「花子は」であり、同様に述語句２は述語句１に依存し、述語句１は主題句に依存するという関係が設定される。このように分解された述語句・主題句は、後述の文選択部１４で依存関係に従って再構成され、要約に取り入れる場合には依存先とまとめた形で取り入れる。図８(b)の文１を例にとると、要約に取り入れられる可能性のあるのは、「太郎は、学校を休んだ」（主題句＋述語句１）あるいは「太郎は、風邪を引いたので学校を休ん

だ」（主題句＋述語句２＋述語句１）のいずれかである。

【００８２】ここで主題句を分離する理由は、後続の文の中に主題句に依存する（主題句について述べている）文が含まれることが多く、そのような文を要約に含める時に主題句のみを含めて可読性を高めることができるからである。例えば、「太郎は学校へ出かけた、途中で犬に出会った。」の第２文「途中で犬に出会った。」を要約に含める時に、省略されている主語を補って「太郎は、…途中で犬に出会った。」（第１文の主題句＋第２文の述語句）として可読性を高めることができる。この場合、主題句を分離しないと、主題句に続く主述語句も一緒に、「太郎は学校へ出かけた。途中で犬に出会った。」全体を要約に含めざるを得なくなる。なお、この処理は、照応表現処理の一貫として後述のステップＳ１５で第２文に対応する述語句「途中で犬に出会った。」から第一文の主題句「太郎は」へ依存関係を設定することで実現される（説明中の実施例ではこの処理は行っていない）。

【００８３】ステップＳ１０で主題句の分離と依存関係の設定が行われた後に、ステップＳ１１で主題句が文の代表句とされ、ステップＳ１３の処理に移行する。ステップＳ９で文に主題句が存在しない場合には、ステップＳ１２で主述語句が文の代表句とされた後に、ステップＳ１３の処理に移行する。ここで代表句とは、文に含まれる主題句および述語句の中で、他の述語句に依存していないものを意味する。すなわち主題句が分離された文については、主題句が代表句であり、それ以外の文については主述語句が代表句である。

【００８４】ステップＳ１３およびＳ１４では、文書構造解析処理で設定されている依存関係が述語句の間の関係に変換される。この処理は文が見出しなどに従属している部分（本文などの依存ブロック）に含まれている場合にのみ実行される。ステップＳ１３で処理中の文が依存ブロック内の要素であるか否かが判定され、要素であると判定されると、ステップＳ１４で処理中の文の代表句と文書構造上の依存先（ブロックの依存先）に対応する主述語句との間に依存関係が設定された後に、ステップＳ１５の処理に移行する。ステップＳ１３で依存ブロック内の要素でないと判定されると、ステップＳ１４の処理を行うことなく、ステップＳ１５の処理に移行する。なおここでは典型的な処理のみを記述しているため、処理中の文より後にある部分に処理中の文が依存するときに依存関係を設定するステップを含んでいない。そのような処理が必要なときには、依存先を指定する条件と従属文の代表句とを記憶しておき、その条件に一致する文を処理する時点で依存関係を設定すればよい。

【００８５】図４の最後のステップ、すなわちステップＳ１５では概念の既知性に基づいた依存関係やペナルティの設定処理が実行される。本実施例では、例えば動詞文の主題語が既知でない場合に、主題語が初めて出現した文を依存先とする依存関係の設定処理と、照応表現を含む文にペナルティを与える処理が実行される。この処理の後に再びステップＳ２の処理に戻り、次の文の取り出しが行われ、ステップＳ３で文が取り出されたと判定されると、ステップＳ４以降の処理が繰り返される。ステップＳ３で文が取り出されなかったと判定された場合には、文分割・依存関係設定処理を終了する。

【００８６】図４の文分割・依存関係設定処理が終了すると、図３の文選択部１４による処理が行われる。この文選択処理は、文分割・依存関係設定部１３から出力される依存関係は述語句列を対象として、要約に含めるべき重要な述語句を選択し、要約に含まれる述語句の選択結果リストを作成するものであり、その処理のフローチャートは図９に示される。

【００８７】図９において、注目情報は注目概念リストとして扱われている。注目概念リストは、具体的には前述のように、注目情報として与えられた重みなしの名詞のリストである。これを注目概念リストと呼ぶのは次の理由からである。本実施例では、各文（述語句）を要約に取り入れるかどうかを判定するのに、文に含まれる注目情報の量を判定の第一の基準として用いており、注目情報量としては、注目情報として与えられた名詞が各々の文にいくつ含まれるかで計算している。注目情報量の計算においては、注目情報として与えられた名詞のリストに含まれる語と字面が一致する文中の語を数えてもよいし、あるいは、「百貨店」と「デパート」、「パソコン」と「パーソナルコンピュータ」のように同じ概念をあらわす語は同一とみなして数えてもよい。この意味で、図９では、注目語リストではなく注目概念リストという言葉を用いている。

【００８８】図９において処理が開始されると、まずステップＳ２０で注目概念リストが作成され、ステップＳ２１で選択結果リストがクリアされる。注目概念リストは、要約プロセス制御部１０および文解析部１２から与えられる名詞リストに文構造解析部で作成者注目の印がつけられた部分に含まれる名詞を加えたものである。要約プロセス制御部１０から与えられる注目情報は、基本的には利用者注目情報であるが、複数の文書の比較のために要約を実行する場合には、比較対象となっている別の文書の作成者注目情報も含まれている。

【００８９】つづいてステップＳ２２で必須出力句リストが空であるか否かが判定され、空でない場合にはステップＳ２３～２４必須出力句を選択結果リストへ加える処理が実行される。ここで、必須出力句とは、利用者が要約プロセス制御部１０を通じて要約に必須で含めるように指示した文書中の要素（見出しなど）と対応する述語句のことである。具体的には、文分解・依存関係設定部１３から出力された述語句のうちで、文書構造解析部１１によって必須出力の印がつけられた部分と対応する

もののことである。ステップＳ２３で、先頭の必須出力句を取り出し、（その先頭の句を必須出力句リストから除いて、）取り出した句を選択結果リストに追加してから、ステップＳ２４で、選択結果リストに追加した句の中の事柄を要約プロセス制御部から送られた既知の事柄のリストに追加し、ステップＳ２２の処理へ戻る。

【００９０】この時選択結果リストに加えた述語句の中に含まれる注目概念を注目概念リストから除くこともできる。但し、通常、必須出力要素として指定されるものは見出しなどであって、要約の核となる概念を含んでいても、完全な文の体裁をとっていないことが多いので、前述の述語句の中の注目概念を注目概念リストから除かない方がよい。述語句が完全な文の体裁をとっている場合に限って、述語句の中の注目概念を注目概念リストから除くという方法が有効である。

【００９１】ステップＳ２２～Ｓ２４の処理が繰り返され、ステップＳ２２で必須出力句リストが空になったと判定されると、ステップＳ２５で文分割・依存関係設定部１３から出力された述語句列のうちで、必須出力句以外の全ての述語句が選択候補とされて、選択候補リストが作成され、続いてステップＳ２６で選択候補リストの中の全ての句に対して注目情報量が計算される。

【００９２】この情報量の計算処理では、典型的にはそれぞれの述語句に含まれる注目概念（名詞）の数が求められる。この時注目情報量を計算しようとしている述語句に依存先があり、依存先がまだ選択結果リストに含まれていない場合には、依存先も含めて注目概念の数を数えることにする。また依存先が複数ある場合には、依存先の注目情報量を先に計算して、注目情報量が最も大きい依存先を選ぶことを原則とする。なお述語句に出現する注目概念の種類（名詞の異なり数）と、延べ数を記憶しておくことが好ましく、また注目概念に重みが与えられている場合には、その重みを乗算して名詞の数を数えるものとする。

【００９３】このように依存先を含めて注目情報量を計算する理由は、前述の概念の既知性基準と関連している。すなわち概念の既知性基準に従えば、例えば文書に固有名詞が繰り返し出現する場合、２番目に出現した部分を要約に含める時には、初めて出現した部分も要約に含めるように処理することが必要となるためである。すなわち文分割・依存関係設定部１３によって、２番目の文から１番目の文に対して依存関係が設定されていることから、文選択部１４では２番目の文の重要度、すなわち注目情報量の計算は１番目の文とまとめて行われることになる。このような処理については具体例を用いて更に後述する。

【００９４】続いてステップＳ２７で注目情報量が０の述語句が選択候補リストから除かれた後に、ステップＳ２８で選択候補リストに残っている全ての述語句に対して新情報量の計算が行われる。新情報量とは、利用者にとって既知ではなく、またすでに選択された述語句には含まれていない事柄に関する情報の量である。この新情報量の計算について図１０の例を用いて説明する。

【００９５】図１０において述語と名詞の組を新情報とし、その組の個数として新情報量を計算する。本文には７個の事柄が含まれるが、そのうち２個は見出しと共通である。見出しを必須出力要素として、図９のステップＳ２３で選択結果リストに追加すれば、本文に含まれる新情報量は５個となる。この例では、述語と名詞の組み合わせを認定するというやや複雑な処理を必要とするが、単純に注目概念（注目語）以外の名詞の数を数えるような、より簡単な方法を用いることも可能である。

【００９６】このように、新情報量の計算としては、概念の列として事柄の情報をモデル化し、すでに選択された述語句には含まれていない事柄の数を数えて新情報量とすればよい。別の方法としては、いわゆる５Ｗ１Ｈのような形式で事柄情報をモデル化し、図５に示したようなフレーム表現によって述語句と比較して、既知の事柄と一致しない述語句の数を新情報量とすることもできる。あるいは５Ｗ１Ｈ形式のモデルを用いた場合の新情報量を第１新情報量とし、単純な新情報量を第２新情報量として、組み合わせて用いることもできる。新情報量の計算も、注目情報量の計算と同様に、依存先があれば新情報量が最も大きな依存先を選んで、依存先を含めて計算するものとする。またペナルティが与えられている述語句に関しては、そのペナルティ分だけ新情報量を減算するものとする。

【００９７】ここでペナルティについて更に説明する。照応表現などに関してペナルティを与える文としては、依存先の文であっても、依存元の文であってもよい。単に新情報量を減算するだけのことである。例えば次の２つの文についてのペナルティを説明する。

【００９８】第１文：昨日[1] 新宿[2] で田中[3] さんに会って、こんな話[4] を聞きました。
第２文：昨日[1] 田中[2] さんに会って、車[3] の話[4] を聞きました。
この例では上に数字の書かれている名詞の数は同じである。しかしながら第１文の方には「こんな話」という照応表現があり、この文だけを取り出すと、どんな話であるかが不明である。照応表現に対するペナルティとは、このような場合に第２文が優先的に選択されるように、第１文に対する新情報量を、例えば名詞の個数で０．５個分減点するものである。

【００９９】簡単な例として、新情報量を文に含まれる注目概念（注目語）以外の名詞の数として計算する。例えば「田中（さん）」が注目語であれば、「こんな話」に対して名詞０．５個分のペナルティを与えると、第１文の新情報量は２．５となり、第２文（新情報量３．０）の方が新情報量が多くなり、後述するように第２文の方が優先的に選択される。しかしこの例で注目語が

「田中（さん）」と「新宿」の２つでれば、第１文の方が注目情報量が多くなるため、後述するように、ペナルティにかかわらず第１文が選択されることになる。

【０１００】ステップＳ２８の処理が終了すると、ステップＳ２９で新情報量が０の述語句が選択候補リストから除かれた後に、ステップＳ３０で選択候補リストが空になったと判定されるまで、ステップＳ３１～Ｓ３６の処理が繰り返される。

【０１０１】まずステップＳ３１で注目情報量が最大の述語句を選び、その中で最大の新情報量を持つ述語句が出力句とされ、ステップＳ３２でその出力句が選択候補リストから除かれて、選択結果リストに追加される。この時述語句に依存先があり、その依存先がまだ選択結果リストに加えられていなければ、その依存先も一緒に選択結果リストに追加する。なお情報量が全く同じ述語句が存在する場合には、それらの述語句を全て一度に追加することを原則とするが、別の方法として最も要約対象文書の先頭に近い述語句を選ぶなど、述語句の出現位置などによって１つに選択してもよい。

【０１０２】このように文選択処理において事柄の既知性に関しては新情報量の差として取り扱われ、注目情報量が同じ文がある場合、新情報量の多い文が選択される。そこで注目情報量が同じ文がなければ、事柄の既知性基準は使用されないことになる。

【０１０３】続いて図９のステップＳ３３で出力句、すなわち選択結果リストに追加された述語句に含まれる注目概念が注目概念リストから除かれ、その結果を用いて選択候補リストに残っている全ての述語句に対する注目情報量の再計算が行われ、ステップＳ３４で再計算された注目情報量が０の述語句が選択候補リストから除かれる。この注目情報量の再計算は前述と同様に行ってもよいが、例えば注目概念と述語句の関係をあらかじめ記憶しておいて、リストから除かれた注目概念を含む述語句と、選択結果リストに追加された述語句に依存している述語句だけを対象に計算を行うこともできる。

【０１０４】ステップＳ３４の処理に続いて、ステップＳ３５で出力句、すなわち選択結果リストに追加された述語句に含まれる事柄の情報が既知事柄リストに追加され、選択候補リストに残っている全ての句に対して新情報量の再計算が行われる。この再計算も前述と同様に行ってもよいが、例えば事柄と述語句の関係をあらかじめ記憶しておき、既知事柄リストに加えられた事柄を含む述語句、選択結果リストに追加された述語句および注目情報量に変化があった述語句に依存する述語句だけを対象に計算を行うこともできる。

【０１０５】図９のステップＳ３６で新情報量が０の述語句が選択候補リストから除かれた後に、ステップＳ３０以降の処理が繰り返され、ステップＳ３０で選択候補リストが空になったと判定された時点で、文選択処理を終了する。

【０１０６】図１１は、前述のように新情報量が第１新情報量と第２新情報量とに区別されている場合の、図９のステップＳ３１における新情報量比較の詳細フローチャートである。同図において、候補述語句Ａと候補述語句Ｂの新情報量を比較する場合には、ステップＳ３８でまず第１新情報量については２つの述語句のうちいずれが大きいかが判定され、Ａの方がＢより大きい時にはＡの新情報量が大きいものと判定され、逆にＢの方がＡより大きい場合にはＢの新情報量が大きいものと判定される。これに対して第１新情報量が同じである場合には、ステップＳ３９で第２新情報量が比較され、第２新情報量の大きい述語句の方の新情報量が大きいものと判定され、第２新情報量が等しい場合には２つの述語句ＡとＢの新情報量は同じと判定される。

【０１０７】以上においてはペナルティは照応表現に対してのみ与えられるものとしたが、前述の依存関係が設定される文（１）～（５）のうちで、（３）と（４）などについてもペナルティを与えておけば、例えば利用者に理解できない用語などの出力が抑制されることになる。その場合の処理は、照応表現についてペナルティを与える場合と全く同様に実行可能である。

【０１０８】あるいは図９のステップＳ３１における新情報量の比較の後で、候補述語句の長さを含めて出力句とするか否かの評価を行うこともできる。すなわち注目情報量と新情報量とが共に等しいものについては、短い述語句を優先的に選択することにすれば、利用者に理解できない用語などの出力はある程度抑制される。更に新情報量そのものの比較に代わって、新情報量と選択述語句の長さとの比（新情報の出現密度）を用いるという方法も考えられる。

【０１０９】いずれにしても、このような利用者に理解できない用語などの出力の抑制の問題については、新情報量の計算とからめて処理することになる。既知概念は基本的には依存関係として取り扱われるが、事柄の既知性基準に基づく新情報量の計算にも関係している。これが、概念の既知性基準と事柄の既知性基準とを、情報の既知性としてまとめた１つの理由である。

【０１１０】以上において本発明における文書要約方式を、一般的に、詳細に説明したが、ここで本発明の特徴について更に説明する。本発明においては、前述のように情報の注目性基準として、利用者注目情報と作成者注目情報の２つが考慮されているために、利用者の求める情報と文書において重要な情報の双方をバランスよく取り込んだ要約を作成することができる。また情報の既知性基準として、概念の既知性基準と事柄の既知性基準とを用いることによって、簡潔で分かりやすい要約が作成されるという特徴がある。

【０１１１】このような前述の特徴に加えて、本実施例によればまず文書の重要性に応じて自動的に要約の長さを変えることが可能となる。従来の要約作成のアルゴリ

ズムにおいては、要約に含めるべき文の数や文字の数、あるいは元の文に対する要約の長さの比率などがパラメータとして渡される場合が多い。本実施例においては、注目情報のうちで文書の中に出現しているものの量に応じた長さの要約が作成されることになり、特別のパラメータを指定することなく、適切な長さの要約が作成される。例えば利用者注目情報を重視すれば、利用者が求める情報に関係の深い文書ほど長い要約が生成されることになる。特に長さにバラつきがある一群の文書をまとめて要約するような場合には、一般に要約の比率などのパラメータを適切に設定することが難しく、この特徴は大きな長所になる。

【０１１２】次に本実施例においては要約の長さに関する制約にも容易に対応できるという特徴がある。本実施例では基本的には要約の長さに関する特別な処理はなされていないが、要約の長さに対して制約が与えられた時には、その制約に容易に対処することが可能である。例えば短い要約を得たい場合には、図９の文選択処理におけるステップＳ３０において、選択候補句リストが空にならない前に処理を途中で打ち切ればよい。これは文選択処理において、注目性基準によって重要度の高い部分から順次文が選択されていることによる。

【０１１３】逆に長い要約を得たい場合には、一旦図９のフローチャートに従って文選択処理を行ってから、選択されなかった部分について図９のフローチャートによる処理を繰り返すことによって、適切な長さの要約を作成することができる。これは図９のステップＳ３１で最初の処理のフローにおいて、注目情報量が最大の句の中で新情報量が最大の句だけが選ばれているために、２回目の処理のフローにおいては次に新情報量が大きい句が選ばれて、出力句とされることによる。すなわち事柄の既知性基準によって、冗長な出力が抑制されているという本発明の特徴がある意味では逆に生かされて、適切な長さの要約の作成に寄与することになる。あるいは、選択処理を繰り返す際に、前回の選択処理で得られた要約中の全名詞を次回の注目情報とするなどという方法でも、効果的に関連性の高い部分を順次取り入れて、要約を拡大することも可能である。

【０１１４】更に本実施例によれば、要約に関する様々な、その他の制約にも容易に対応できるという特徴がある。本実施例においては要約の満たすべき要件を、情報の注目性基準および既知性基準という２つの基準に対応して整理して用いることができ、いろいろな要求に応じて文書要約装置の動作を拡張することが容易である。例えば図３において、利用者の嗜好特性や利用者の知識のようなメモリに格納されている情報を、利用者の要求に応じて様々な面から再整理して、要約作成の制約として用いることができる。また、２つの文書の作成者注目情報を、互いに別の文書を要約するための利用者注目情報のような形で与えて、要約を作成すれば、それぞれの文書で共通して述べられている事柄や片方にしか述べられていない事柄のうち、どちらかの文書の作成者が重点を置いているものが抽出されることになるので、そのような要約を文書の比較情報として用いることも可能である。このように要約装置の基本構成を変えることなく、各種の要求に対応することが可能である。

【０１１５】続いて本発明の文書要約方式を用いた要約作成の具体例について説明する。図１２は本明細書の〔従来の技術〕などで参照した特許公報の抄録を対象として、次の検索式と見出し（発明の名称）を注目情報として、要約としての抄録抜粋を作成した結果を示す。

【０１１６】（要約ＯＲ抄録ＯＲ読解ＯＲ閲覧）ＡＮＤ（文書ＯＲドキュメント）

なお最後の特開平７－４４５６６だけに対しては、検索式は以下のものである。

【０１１７】（要約ＯＲ抄録ＯＲ読解ＯＲ閲覧）ＡＮＤ（文章ＯＲテキスト）

図１２において、要約としての抄録抜粋の中の〔　　〕でくくって表示されているものは注目語である。また“特徴語”は要約対象文書の中に含まれていた注目情報である。ここで特徴語は注目語の部分集合になるが、機能的には別のものである。特に利用者が注目している語の中で、ある文書に出現したものは、その文書に対するキーワードには含まれていなくても、利用者にとっては意味が深い、その文書の特徴と考えられる。

【０１１８】図１２おける要約作成の条件は❶利用者注目情報として検索式に含まれている名詞を用いる。❷作成者注目情報として見出し（発明の名称）に含まれている名詞を用いる。❸概念の既知性基準は用いない。❹既知の事柄情報としては選択された要約部分に含まれている名詞を用いる。すなわち候補述語句中に存在し、まだ要約に含まれていない名詞の数が新情報量（異なり数が第１新情報量、延べ数が第２新情報量）とされる。❺見出し（発明の名称）を必須出力要素とする。の５つである。

【０１１９】図１３は概念の既知性基準を用いた場合の効果を説明する要約作成の具体例である。これは経済関係のレポートに対して、見出しを注目情報として、要約を作成した例である。図１３(a)は「Hancock は」という主題句に関する概念の既知性基準を用いない場合の要約であり、(b)はこの主題句に関する概念の既知性基準を用いた場合の要約である。概念の既知性基準によって追加された部分にはアンダーラインがつけられている。

【０１２０】図１３における要約作成の条件は❶利用者注目情報は指定しない。❷作成者注目情報として見出しに含まれる名詞を用いる。❸概念の既知性基準は(a)では使用せず、(b)では主題句に関して用いる。❹既知の事柄情報としては選択された要約に含まれている名詞を用いる。すなわち候補述語句中に存在しながら、まだ要約に含まれていない名詞の数が新情報量（異なり数、延

べ数が第１、第２新情報量）とされる。⑤見出しを必須出力要素とする。の５つである。

【０１２１】図１３の要約作成処理について更に詳細に説明する。図１３における要約対象文書は以下のものである。この文書において、下線の付いた文字で以下の記号のついた文が、図１３で要約として抜粋された文である。

【０１２２】・図１３(a) でも抽出されている文（◇）
・図１３(b) で新たに追加された文（☆）

○ Apple ComputerがWindows 乗り入れの強化により再建中

◇ G.Amelioは社内の機構を改革し、Macintosh の機種を半分に減らして開発費を減らし、約3,000 人の社員をレイオフしつつ、Apple Computerを建て直している。

【０１２３】☆ AmelioはApple の重要な地位に外部から人を入れているが、研究開発の最高の担当者chief techhology officerとして５３才のEllen Hancock を任命した。これはApple 再建に最も重要な地位である。これは業界で非常に尊敬されていたD.Nagel がAT&TのBell Laboratories の所長としてApple を去るまで占めていた地位である。有意義な新製品開発の経験のある人の代わりにＩＢＭに28年間過ごしたHancock が任命されたのは驚きを持って迎えられている。Hancock はＩＢＭという巨大な会社で育っているので、直ちに6,000 人の血気盛んな若いエンジニアやプログラマと管理スタイルの上でぶかつり合うのではないかと見られている。また、ＩＢＭでは５年位の単位で動いているのに対し、Apple は直ちに動かなければならないから、仕事のテンポが合わないのではないかとの懸念もある。Hancock は数学の修士を持ち、１９９６年にＩＢＭのプログラマーとして出発し、管理能力が認められて次第に昇格し、１９９５年にはＩＢＭ全体の約1/3 を担当したが、L.Gerstnerと意見が合わず、ＩＢＭを去ってNational SemiconductorにＣＯＯとして迎えられた。Hancock はＩＢＭはLotus Development のNotes を買い取るべきだと長いことＩＢＭ社内で説いていたが、それが実現したのはHancock がＩＢＭを辞めてからである。National Semiconductorでは今年２月にAmelioがApple に移った後、後任のchief executive officer になるつもりでいたが、board of directorsが後任にＬＳＩ LogicからのB.Halla を任命したので、National Semiconductorを辞めた。◇しかし、Hancock はソフトウェアをよく知っており、Apple 再建の成否は開発が遅れ続けているCopland にかかっているので、Hancock は妥当な人事と見られている。また、Apple では開発管理がいい加減で製品化が不首尾になることが多かったのを是正し、また今まで大企業のマーケットに進出できなかったのを是正できるのではないかと期待されている。

【０１２４】その他chief operating officer としてTexas Instruments からはMarco Landi, chief administrative officer としてはMaxtor Corp., Advanced Micro Devices, Fairchild Semiconductorなどを経たGeorge H.Scalise、またchief financial officer としてAutomatic Data Processing Inc.とMAI Systems にいたことのあるFred D.Anderson をAmelioは任命している。

【０１２５】Amelioの前任者Spindlerの時はSpindlerの無理な開発促進圧力のため、開発の中心人物が多数Apple を辞め出した。しかもその多くがMicrosoft に就職している。S.Capps は１５年間Apple にいてMacintosh その他のヒット商品を考え出したが、６ケ月前からApple を去る決心をし、新会社を興すためベンチャー・キャピタリストと語ったところアイデアが多過ぎると言われたため、諦めてApple の競争相手のMicrosoft に就職し、Microsoft のInternetのツールと新しいcomputer interfaceを開発し出している。Microsoft は今年末までにNewtonのようなhand hold computer用ＯＳのPegasus を発表する予定だが、それを使い易くするのに協力する。またCapps と共にNewtonを開発したW.Smith もMicrosoft に移った。Gates がもっと使い易いインターフェイスを求めているからその方針に従うが、Windows は二人にとっては初めての経験である。Windows95 は複雑で同じことをするのに五つもの異なるやり方があり、単純化するのは困難と見られている。

【０１２６】AmelioはApple 再建に妥当な手を打っていると見られるが、結果が目に見えてくるまでは少なく共１年はかかるであろう。しかし、Macintosh の売れ行きは悪化している。今年の３月に終わる四半期のApple の売上は１年前に比べ９．７％下がって＄２．８billion になった。Macintosh の出荷は６月に終わる四半期には２０％減ると業界では推定していたが、調査会社Computer Intelligence が1,000 のパソコン小売店を調べたところ、アメリカでは４月と５月における出荷台数はそれより遙かに悪く、１年前に比べて４月は２９％、５月には２７％減り、売上金額は４月には３１％、５月には３３％減っているという。この減少の一部にはパソコン業界全体の売れ行き鈍化と、春先に欠陥のあったMacintosh を多数リコールしなければならなかったのも含まれている。しかし、Merisel Inc.のようにMacintosh は今まで通り売れているとい所もある。パソコン業界全体では４月も５月も売上は１０％増え、出荷台数は３％増えている。大きな減少は企業のマーケットであり、社内のパソコンの半分から1/4 までがMacintosh の3,000 のオフィスを対象に調べたところ、新規のパソコンの購入が２月に３３％あったのに４月には１４％減っている。アメリカ最大のパソコン量販チェーンのCompu USA Inc.では、Macintosh の売上は５０％も下がったが、ノートブックの売れ行きはいくつかの機種がリコールされたこともあってストップしている。小売店ではどこでもMacintosh のハードウェアもソフトウェアも余り置いていないが、それは通信販売の会社から安く買えるためである。

そういう通信販売会社の最大がMicro Warehouse であり、毎年$1.8billionの売上があるが、その半分がMacintosh のハードウェアとソフトウェアであり、夜10時までに電話やFAX で注文すれば翌日＄３の送料で配達するという優れたサービスで有名である。この会社では１月におけるMacintosh の売上は１年前に比べ６０％増えたが、５月には増減なしとなった。

【０１２７】Dataquest によれば、世界のマルチメディアのマーケットではMacintosh が最大で、１９９５年には3,950,000 台（１９９４年には2,400,000 台）、次がPackard Bell で3,000,000 台（2,950,000 台）、それに続いてCompaqが2,900,000台（1,200,000 台）、ＩＢＭが1,600,000 台（800,000 台）、ＮＥＣが1,500,000 台（500,000 台）と続く。１９９５年のマーケット・シェアはApple が最大で、２２．９％、これに続いてPackard Bellが１９．２％、Compaqが１１．９％、ＩＢＭが８％、ＮＥＣが４．３％、Acerが２．７％、Escom が０．７％、富士通が０．６％、Highscreenが０．６％、その他が２９．１％となっている。

【０１２８】Apple は５月にdigital cameraやその他の画像処理用装置に内蔵されるチップ上で動作する新しいＯＳを発表した。これはQuick Time IC(image-capture)技術の一部であり、MotorolaのチップＭＰＣ８２３用のmultitasking OS であり、image-capture 用装置のＡＰＩを含んでいる。現在digital cameraの製造会社はそれぞれ独自のＡＳＩＣを設計し、Adobe のPhtoshopやStorm Software社のEasyPhoto といった画像処理用ソフトウェアのインタフェースを独自に開発しなければならない。QuickTime ICを使えばdigital cameraの製造会社はそういう手間が省けるので、digital camreaの値段を下げられる。Apple はこれの開発をdigitalcameraや画像処理関係の会社の大手と共同で開発してきており、すでに10社以上が支持している。これを使えばパソコンなしにdigital cameraから直接Internetに画像を送れるし、カメラの中のscriptにより撮影時間を変えたり、Photoshopのフィルターを動作さすことができる。

【０１２９】Apple はMacintosh 互換機戦略を強化しつつあるが、そのための製品を夏から出荷する。社内でTanzaniaと呼ばれる新しいMacintosh のlogic board はscalableで安く、これによってMacintosh 互換機の製造に興味を持つ会社に呼びかける。Motorolaはすでに Tanzaniaの試作を済ませ、実演をした。Tanzaniaは低位と中位機種用であり、広範なオプションがある。最高200MHzまでのPowerPC603e と604eを使用出来、３個から５個までのＰＣＩスロットがある。ＰＳ／２キーボードかＡＤＢコネクター、またEnhanced IDE (Integrated Drive Electronic)かＳＣＳＩ internal hard driveの選択がある。またLocalTalk, GeoPort, ＳＣＳＩなどのコネクタの他、Apple としては初めてのＡＴＡＤＩ (AT attachment packetinterface) ＣＤ－ＲＯＭ driveがある。TanzaniaはまたMacintosh の自動ejectingのfloppy driveの他、Intel のチップ使用のパソコンでは標準の手動ejectingのものもある。ＤＩＭＭスロットが２個とＳＩＭＭのスロットが二つあり、最大160Mbytes のＲＡＭが使えるがＥＤＯ ＤＲＡＭを使用する。互換機製造会社はTanzaniaを使用したパソコンを来年始めには出荷できる。Apple はMacintoshのライセンス戦略をMacintosh そのものの互換だけからＰＰＣＰへの移行を３段階に分けて推進している。第１段階はMacintosh そのものの互換機だけであって１９９５年から１９９６年にかけてDayStar Digital Inc.、 Power Computing Corp.、 Umax Computer Corp.がPower Macintosh7500 と9500の互換機を実現した。Umax Computer Corp. は台湾のUmax Data Systems がRadius Inc. からMacintosh 互換機部門を買い取って今年１月に生まれた会社であ、その最初の互換機SuperMacS900は６月始めから出荷されたが、非常な人気で生産が間に合わず、１カ月以内に＄10million の受注がこなせないでいる。第二段階はPower Macintosh5400 とTanzaniaボードに基づいて今年夏から来年半ばにかけて実現する。この二つは共にLow End Reference Platform（最近はMacOS Licensing Design，略してＭＬＤという）に基づいている。どちらも業界で標準になっている広範囲の論理回路や周辺機器が使えるようになっており、ＰＰＣＰ（以前はCommon HardwareReference Platform 、略してＣＨＲＰと呼ばれていた）の狙いに近づく。ＰＰＣＰではMacintosh のＯＳの他、OS/2, Windows 3.1, UNIX, Solarisなど広範囲のＯＳが使えることになっている。第三段階はＰＰＣＰに１９９７年半ばから１９９８年にかけて完全に移行する。

【０１３０】Microkernel に基づくCopland が来年半ばに延びたので、今までApple はCopland が出るまでSystem7.5.3 がSystemの改良の最後だと言っていたのを変更して、つなぎとしてCopland の新しい機能の一部を取り入れたHarmony と社内で呼んでいるＯＳを年末に発表することになった。Harmony にはInternetへのサポート、OpenDoc, Cyberdog, QuickTime2.5, QuickDraw3Dなどのクラフィックス技術、またCopland に予定されていたインターフェイスの改良などがある。またラベルの付いたフォルダーによってファイル多数の検索や管理もできる。Lockheed Martin Missiles and Space 社はMacintosh を9,500 台持っているが、Internetへのサポートや OpenDoc が完成しているのなら、来年まで待たずに入手できるのは有難いと大歓迎である。

【０１３１】ユーザは今までのソフトウェアを変更せずにHramony を使えるが、Copland の場合はソフトウェア会社は今までのソフトウェアを変更しなければならない。Copland は今はSystem8 と呼ばれている。

【０１３２】また今年の夏にはSystem7.5.3 のバグを直し、Duo23005やPowerBook に対する性能を改善し、Busterと社内で呼ばれて開発されてきたものを発表する。Co

mputer Intelligence InfoCorp.の最近の調査では、昨年Macintosh を買った人の中で次もMacintosh を買おうと答えたのが８７％もあり、パソコンの満足度では最高であった。次いでDell Computer が７４％、Hewlett-Packard が７２％、Acerが６８％、Gateway2000 が６１％であった。Macintosh に満足している人々はMacintoshのＯＳが好きなためであるのに対し、Intel チップ使用のパソコンを使用している人々はたとえMacintosh のＯＳの方がすきであってもソフトウェアの互換性の立場からMacintosh に変えられないとこの調査会社は説明している。

【０１３３】Amelioの前任者SpindlerはMicrosoft を徹底的に敵視してGates に会うことはなかったが、AmelioはGates を訪問して協力を要請するという現実的な行動をしている。マルチメディアに関する標準と製品をApple とMicrosoft の二社で共通に使用しようとしている。二社の交渉がまとまればApple のQuickTime の開発環境がWindows95 やDirectX API のサポートも含めＮＴでも使えることになり、またApple のQuickTime Internet Explorer がMicrosoft のInternet Explorer でも使える。◇交渉がまとまればMicrosoft はQuickTime をInternet Explorer に組み込めるようにし、同時にApple はWindows のマルチメディア技術のサポートを強化することになろう。すでにApple はQickTimeがwindows 使用の環境でも使用出来るようにし、DirectX のＡＰＩの多くがQuickTime でも使えるようになっている。今まではWindows 用のQuickTime は再生しかできなかった。Web 上のビデオの６０％がOpenDoc で作られ、３０％がＭＰＥＧにより作られるようになりつつあり、ＭＰＥＧはQuickTime でも読めるという事実の前に、Microsoft は現実的になってきた。

【０１３４】二社の関係は他の点でもよくなってきている。Apple はMicrosoft のBackOfficeをヨーロッパではAdvanced Workgroup Solutionsのサーバーにバンドルして売り出しており、これが成功すればアメリカでもそうすると言う。Microsoft もOffice97の次のversion をMacintosh 用にも開発すると約束するようにまで二社の関係は改善された。パソコンの需要が落ち込みつつある現在は特に他社のソフトウェアやハードウェアへの相互乗り入れは互いに利益がある。

【０１３５】Apple はQuickTime Internet技術をMicrosoft にライセンスを与えるばかりでなく、QuickTime VR (virtual reality) も含めることになろう。これに対抗するMicrosoft のActiveMovie の技術は非常に遅れており、開発キットをソフトウェア会社に配布さえしていない。Microsoft はActiveMovie をInternet Explorerに取り入れるのを諦めるのではないかとみられている。しかしApple はDirectXのＡＰＩなどMicrosoft のメディア技術のサポートを改善するためにQuickTimeを貫き直さなければならない。DirectX のＡＰＩの大部分をApple はサポートしつつある。これらＡＰＩのうちDirect3Dは、Apple のQuickDraw3D と真っ向から対抗する。他方Intel もInternet上のビデオ技術についてMicrosoft に働きかけている。Apple のビデオ技術よりIntel のビデオ技術の改良されたものの方がよいとMicrosoft に説いている。Apple の場合はビデオを再生し始める前に充分なデータをダウンロードしなければならないのに対し、Intel の改良した技術では圧縮技術の改良によって早くビデオを再生し始められる。Gates はInternetやintranetの将来性を見誤ったのに気がつき、今は出遅れたInternet Explorer を一刻も早く強力なものにしたいとあせっている。

（要約対象文書終わり）この文書を対象とする要約の作成における注目情報としては、見出しに含まれている名詞を用いる。すなわち注目語は“Apple Computer”、“Windows ”、“強化”、および“再建”の４つである。注目情報量についても、第１注目情報量と第２注目情報量とに区分して取り扱う。第１注目情報量は文に含まれる注目語の異なり数であり、第２注目情報量は文に含まれる注目語の延べ数である。第１注目情報量、および第２注目情報量の取り扱いは、図１１の新情報量の比較と同様に行われる。

【０１３６】説明の単純化のために、新情報量としては注目語以外の名詞（内容語）の数を用いることとし、第１新情報量は文に含まれる注目語以外の名詞内容語の異なり数、第２新情報量は文に含まれる注目語以外の名詞内容語の延べ数とする。

【０１３７】図１４は注目情報量と新情報量の計算結果である。上の要約対象文書の先頭から文に番号をつけ、注目情報量、新情報量が０とならない文についての情報量を示す。文に対して注目語は太字の〔 〕、新情報としての注目語以外の名詞内容語は細字の〔 〕で囲んで表示されている。情報量の少数点より前の部分が異なり数であり、少数点より後の部分が延べ数である。例えば文１１においては〔Hancock 〕が２回出現するので、新情報量の異なり数（第１新情報量）は８、延べ数（第２新情報量）は９となる。

【０１３８】図１４の計算結果に基づいて、図９のステップＳ３１で文７２が選択され、ステップＳ３２で文７２が選択候補リストから除かれ、選択結果リストに追加される。そしてステップＳ３３で“Windows ”と“強化”とが注目語リストから除かれ、注目情報量の再計算が行われる。

【０１３９】図１５は再計算後の情報量を示す。この図に対しては、図９のステップＳ３１で文１が選択され、ステップＳ３２で選択結果リストに追加され、ステップＳ３３で“Apple Computer”が注目語リストから除かれ、注目情報量の再計算が行われる。ここでは文１以外に“Apple Computer”を含む文は存在せず、他の文の情報量は変化しない。

【０１４０】続いて再びステップＳ３１で文１１が選択され、ステップＳ３３で“再建”が注目語リストから除かれ、注目語リストが空になる。従って、注目情報量を再計算すると、選択候補リストに残っている述語句に対する注目情報量は全て０になり、ステップＳ３４で選択候補リストの内容は空となり、文選択処理が終了する。この処理によって得られた結果が図１３(a) である。

【０１４１】次に図１３(b) の結果を得る処理について説明する。ここでは図１３(a) に対する処理に加えて、以下の処理が実行される。まず第１に、主題句の中に既知でない固有名詞が表れた文に対しては、その固有名詞が要約対象文書中で初めて出現した文を依存先とする依存関係を設定する。但し固有名詞の場合、最初は正式名称（ここでは“Ellen Hancock ”，“G.Amelio”）が用いられても、その後は省略形（ここでは“Hancock ”，“Amelio”）が用いられることが多いので、そのような正式名称と省略形は同じとみなす。第２に主題句に指示詞（例えば「これ」）が出現した時には、その直前の文を依存先として依存関係を設定する。第３に依存先の文が、更に第１および第２の処理における依存関係に該当する場合には、その先の依存先の文についても同様に依存関係を設定する。

【０１４２】まず注目語が出現する文に対して第１～第３の処理を行って、依存関係を設定する。図１６はこの依存関係を示す。例えば文１１の主題句としての「Hancockは」に関して文１１から文２に対する依存関係が設定され、文２の「Amelio」に関して文２ら文１への依存関係が設定されている。なお文４１や文７２における“Apple ”や“Microsoft ”も固有名詞であるが、有名な企業名であり、利用者既知概念として与えられたものとして、以下の説明を行う。

【０１４３】このような依存関係を考慮した情報量の計算について、文１１を例にとって説明する。図１６(c) の依存関係に対応して、文１１の情報量は文２と文１の情報量を含めて計算する。ここで“Apple ”は“Apple Computer”と同一の語とみなす。

【０１４４】図１７はその計算結果を示す。文１１の注目情報量は文１１に“再建”、文１に“Apple Computer”が含まれているため、異なり数としての第１注目情報量は２、延べ数としての第２注目情報量は“Apple ”の分も含めて４になる。また新情報量は“Apple ”を除き、“Hancock ”，“Ellen Hancock ”，“G.Amelio”，“Amelio”、および“開発”の重複を除いて、異なり数で２４、延べ数で２７となる。

【０１４５】図１７の計算結果を用いて、図９のステップＳ３１で文１１が選択され、Ｓ３２で文１および文２と共に選択候補リストから除かれ、選択結果リストに追加される。これによって、例えば文３の依存先は選択済となるので、次に文３が選択される場合には、その情報量は文３だけに対して計算される。その後ステップＳ３３で“Apple Computer”と“再建”が注目語リストから除かれ、注目情報量の計算が行われ、文３と文２１の注目情報量が０となる。この結果が図１８である。

【０１４６】更に図１８の結果によって、ステップＳ３１で文７２が選択され、選択結果リストに追加され、ステップＳ３３で“Windows ”と“強化”とが注目語リストから除かれ、注目語リストが空になるために、文選択処理が終了する。これによって図１３(b) の結果が得られる。

【０１４７】最後に本発明の文選択方式の第２の実施例について説明する。図１９は、この文選択方式としての文抽出のアルゴリズムを示す。このアルゴリズムは、新聞記事やレポートなどの見出しに含まれる名詞キーワードを用いて、その名詞キーワードを含む文を抽出して、記事などのダイジェスト情報を作成するアルゴリズムである。

【０１４８】図２０は、図１９のアルゴリズムにおける用語と、図９の文選択処理のフローチャートにおける用語との対応の説明図である。なお図９に対しては、図１４に関する説明と同様に、注目情報量が第１情報量と第２情報量とに区別されている。

【０１４９】例えば図１２と比較すると、図１２においては検索式（質問文）が用いられているのに対し、図１９では見出しだけが用いられる点が異なっている。見出しと質問文の違いとしては、まず第１に見出しは必須出力要素であり、見出しと全く同じ語しか含まない文、すなわち新情報量が０の文は抽出されないことと、第２に質問文（検索式）は単なる注目語のリストであり、質問文と全く同じ語しか含まない文であっても抽出される、すなわち質問文自体は選択結果リストに含まれず、新情報量が０にならないということがある。

【０１５０】図２０において、新情報量と見出しキーワードに一致しない名詞の延べ数とを対応させることは、注目概念（見出しキーワード）と同じ文に出てくる名詞の組の数によって新しい事柄の量を求めるという考え方と同じである。すなわち（３）の比較、注目概念（正確には選択結果リストにまだ含まれていない注目概念）の数が全く同一の文について行われるため、まだ出現していない注目概念とそれ以外の名詞の組み合わせを数えることと同じになる。

【０１５１】図１９においては見出しだけを注目概念のソースとして用いるために、注目概念同士の組は見出しにすでに出現していることになり、図９では必須出力要素としてすでに選択結果リストに含まれていることになる。それ以外の注目概念に関する名詞の組の数は、文に含まれる見出しキーワードの数と見出しキーワード以外の名詞の数との積で与えられ、見出しキーワードの数が同じとすれば、見出しキーワード以外の名詞の数だけを（３）において比較することにより、名詞の組の数の比較を行うことと同等となる。

37

【０１５２】

【発明の効果】以上詳細に説明したように、本発明の文書要約装置を用いることにより、前述の様々な特徴によって各種の効果が生ずるが、最も大きな第１の効果として、文書の有用性（関連度）の判定が容易になるという効果がある。すなわち本発明の方法によれば、利用者が注目している情報と、作成者が重点的に記述しようとしている情報との両方を要約の中に抽出することができるため、利用者が注目している情報が文書の中でどのように扱われているかが、要約を読むだけで容易に把握可能となる。すなわち、文書が利用者の目的とどの位関係があるかということを、要約から容易に判定できるようになる。

【０１５３】第２の大きな効果としては、要約の可読性が向上するという効果がある。概念の既知性基準によって利用者が知らない用語については、例えば追加説明と共に出力され、また事柄の既知性基準によって冗長な出力が抑制されるので、簡潔かつ分かりやすい要約が作成される。また利用者注目情報に基づいて、利用者の注意の方向を考慮することにより、利用者に不必要な情報が要約中に含まれる率を減少させることができることも、利用者にとってより分かりやすくする上で大きな効果がある。

【図面の簡単な説明】

【図１】本発明の第１の原理の説明図である。

【図２】本発明の第２の原理の説明図である。

【図３】本発明の文書要約装置の構成を示すブロック図である。

【図４】文分割・依存関係設定処理の詳細フローチャートである。

【図５】格フレームによる事柄情報の表現の例を示す図である。

【図６】意味ネットワークによる事柄情報の表現の例を示す図である。

【図７】照応表現にペナルティを与える理由を説明するための例を示す図である。

38

【図８】述語句への分割と依存関係の例を示す図である。

【図９】文選択処理の詳細フローチャートである。

【図１０】新情報量の計算方法を説明する図である。

【図１１】第１と第２の新情報量を区別する場合の新情報量の比較処理フローチャートである。

【図１２】特許抄録の要約結果の例を示す図である。

【図１３】主題句に関する概念の既知性基準の使用の効果を説明する要約例を示す図である。

【図１４】図１３(a) を得るための情報量の初回計算結果を示す図である。

【図１５】図１４において文７２が選択された後の情報量を示す図である。

【図１６】図１３(b) の結果を得るための依存関係の設定を説明する図である。

【図１７】図１６の依存関係を考慮した情報量の計算結果を示す図である。

【図１８】図１７において文１１が選択された後の情報量を示す図である。

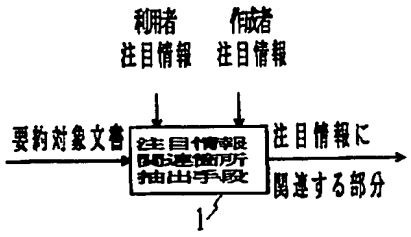
【図１９】文選択方式の他の実施例としてのダイジェスト情報抽出のアルゴリズムを示す図である。

【図２０】図１９のアルゴリズムと図９のフローチャートとにおける用語の対応を示す図である。

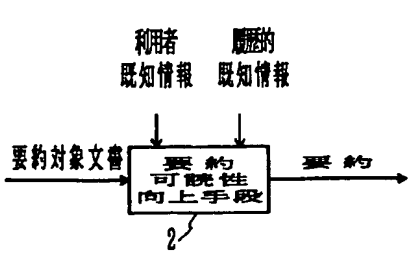
【符号の説明】

１　注目情報関連箇所抽出手段
２　要約可読性向上手段
１０　要約プロセス制御部
１１　文書構造解析部
１２　文解析部
１３　文分割・依存関係設定部
１４　文選択部
１５　要約整形部
１６　利用者の嗜好特性
１７　利用者の知識
１８　閲覧履歴
１９　入力文書（群）

【図１】

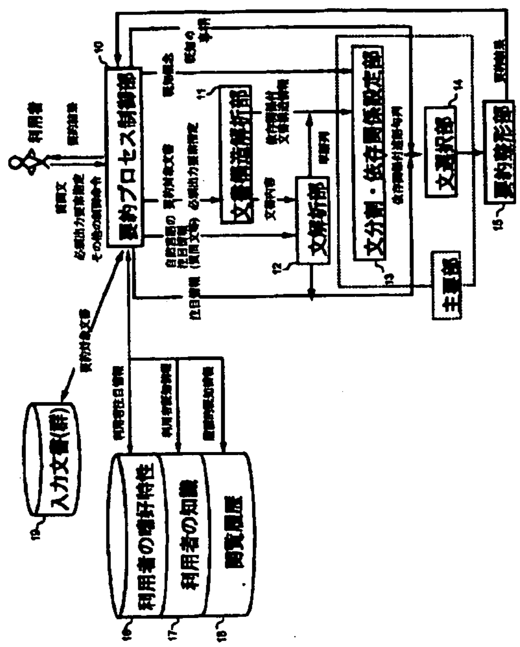


【図２】



【図３】

本発明の文書要約装置の構成を示すブロック図



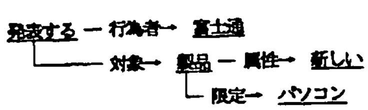
【図５】

格フレームによる事柄情報の表現の例を示す図

動詞:　発表
ガ格:　富士通
ヲ格:　パソコンの新製品
時:　平成３年３月３日
場所:　日本

【図６】

意味ネットワークによる
事柄情報の表現の例を示す図



【図７】

照応表現にペナルティを与える
理由を説明するための例を示す図

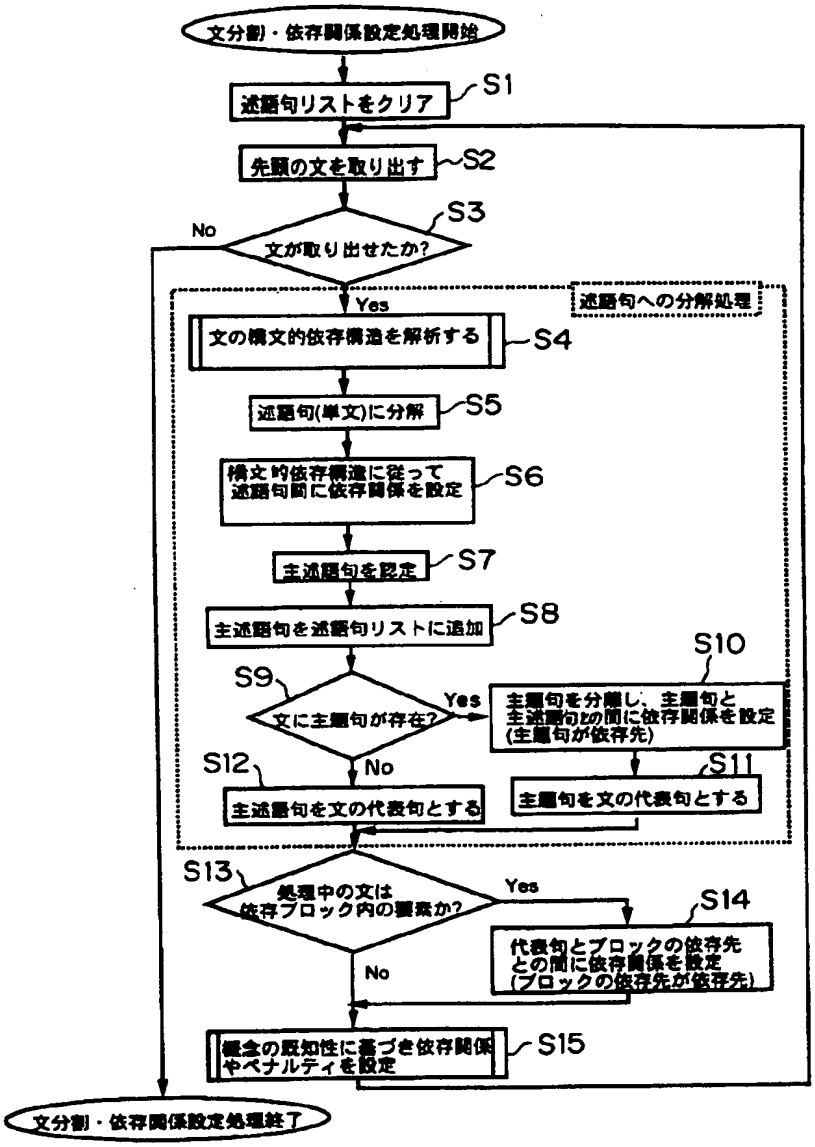
…この企業ＯＳトリオは、絶滅の危機にはない。調査対象となったほとんどのサイトで、いまだにハイエンドではＯＳ／３９０、ミッドレベルでＵＮＩＸ、そしてデパートメンタルＬＡＮベース・レベルではＮＴが使われている。

…日本の米パソコン市場への進出は、これまでポータブルに限られていた。だが、コンピュータのマルチメディア化が進み、テレビとの融合が進むにつれて、日本企業は有利になるかもしれない。この段階は、日本のライバルにとっては微妙だ。ＣｏｍｐａｑＣｏｍｐｕｔｅｒは日本の最新のチャレンジについて、コメントを辞退した。…

（…は段落境界以外で省略した文がある部分を示す。）

【図４】

文分割・依存関係設定処理の詳細フローチャート



【図８】

述語句への分割と依存関係の例を示す図

文１：　太郎は、風邪を引いたので学校を休んだ。
述語句１：　太郎は、／学校を／休んだ。
述語句２：　風邪を／引いたので
依存関係：　<述語句１>←依存－<述語句２>

文２：　花子は、太郎が送ってくれた手紙を大事そうにしまった。
述語句１：　花子は、／手紙を／大事そうに／しまった。
述語句２：　太郎が／送ってくれた
依存関係：　<述語句１>←依存－<述語句２>
(a) 主題句分離前

文１：　太郎は、風邪を引いたので学校を休んだ。
主題句（代表）：　太郎は、
述語句１（主）：　学校を／休んだ。
述語句２：　風邪を／引いたので
依存関係：　<主題句>←依存－<述語句１>←依存－<述語句２>

文２：　花子は、太郎が送ってくれた手紙を大事そうにしまった。
主題句（代表）：　花子は、
述語句１（主）：　手紙を／大事そうに／しまった。
述語句２：　太郎が／送ってくれた
依存関係：　<主題句>←依存－<述語句１>←依存－<述語句２>
(b) 主題句分離後

【図１３】

主題句に関する概念の既知性基準の使用の効果を説明する要約例を示す図

Title: Apple Computer が Windows 乗り入れの強化により再意中

G.Amelio は社内の機構を改革し、Macintosh の機器を予分に絞らして開発費を絞らし、約 3,000 人の社員をレイオフしつつ、Apple Computer を建て直している。…しかし Hancock はソフトウェアをよく知っており、Apple 再建の成否は開発が遅れ続けている Copland にかかっているので、Hancock は妥当な人事と見られている。…交渉がまとまれば Microsoft は QuickTime を Internet Explorer に組み込めるようにし、同時に Apple は Windows のマルチメディア技術のサポートを強化することになろう。…

(a) 概念の既知基準を用いない場合

Title: Apple Computer が Windows 乗り入れの強化により再意中

G.Amelio は社内の機構を改革し、Macintosh の機器を予分に絞らして開発費を絞らし、約 3,000 人の社員をレイオフしつつ、Apple Computer を建て直している。Amelio 社 Apple の重要な地位に外部から人を入れているが、研究開発の最高の責任者 chief technology officer として 53 才の Ellen Hancock を任命した。…しかし Hancock はソフトウェアをよく知っており、Apple 再建の成否は開発が遅れ続けている Copland にかかっているので、Hancock は妥当な人事と見られている。…交渉がまとまれば Microsoft は QuickTime を Internet Explorer に組み込めるようにし、同時に Apple は Windows のマルチメディア技術のサポートを強化することになろう。…

(b) 主題句に関して概念の既知性基準を用いた場合

【図１０】

新情報量の計算方法を説明する図

見出し：　富士通新型情報機器発売へ

本文：　富士通は三十日、新型情報機器の発売を三月に開始すると発表した。

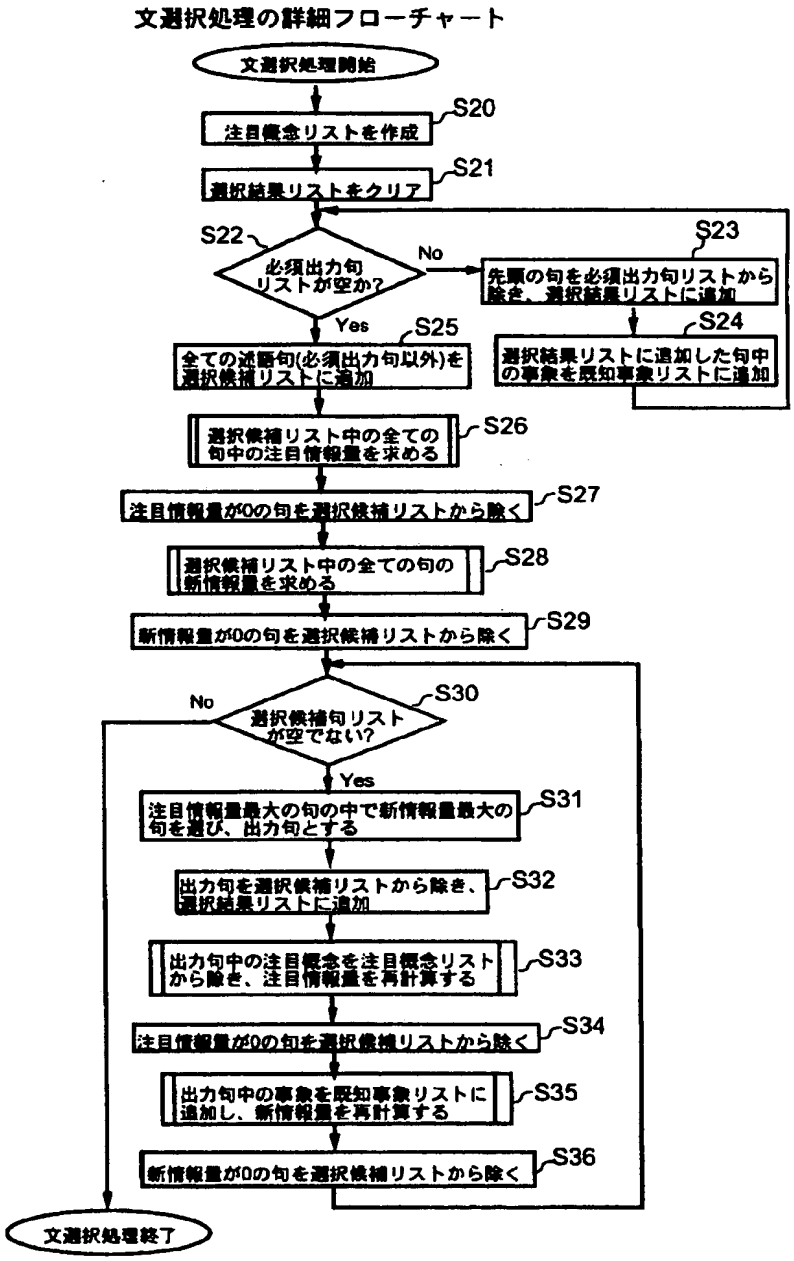
見出しに含まれる事例	本文に含まれる事例
(富士通, 発表) (新型情報機器, 発表)	(富士通, 発表) (新型情報機器, 発表) (発表, 開始) (三月, 開始) (富士通, 発表) (三十日, 発表) (開始, 発表)

【図１４】

図 13 (c) を得るための情報量の初回計算結果を示す図

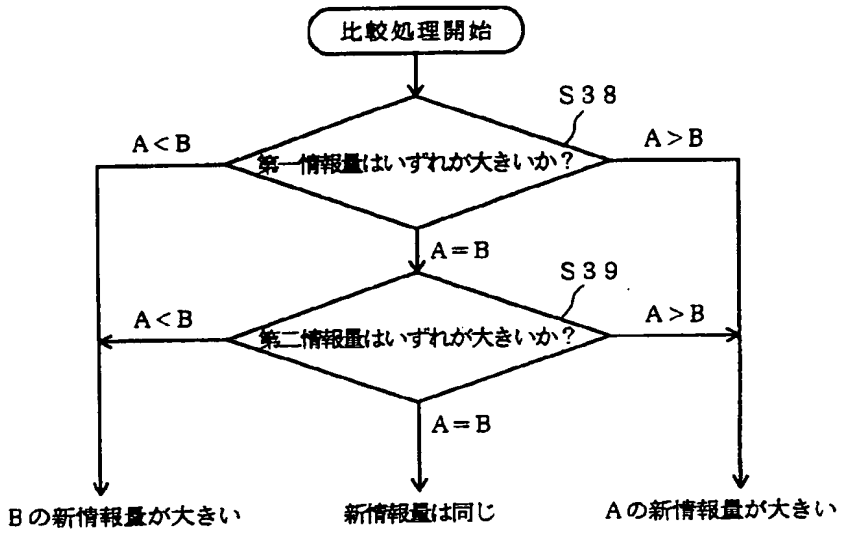
文	新情報量	注目情報量	文番号
(G. Amelio) は (社内) の (機構) を (改革) し、(Macintosh) の (機器) を予分に絞らして (開発) 費を絞らし、約3,000 人の (社員) を (レイオフ) しつつ、(Apple Computer) を建て直している。	0.0	1.1	1
これは (Apple) (再建) に (重要) な (地位) である。	3.3	1.1	9
しかし (Hancock) は (ソフトウェア) をよく知っており、(Apple) (再建) の (成否) は (開発) が遅れ続けている (Copland) にかかっているので、(Hancock) は (妥当) な (人事) と見られている。	0.0	1.1	11
(Amelio) は (Apple) (再建) に (重要) な (手) を打っていると見られるが、結果が (目) に見えてくるまでは少なくとも1年はかかるであろう。	5.5	1.1	21
(Apple) は (Macintosh) (互換) 機 (機器) を (強化) しつつあるが、そのための (製品) を更から (出荷) する。	0.0	1.1	41
(交渉) がまとまれば (Macintosh) は (QuickTime) を (Internet Explorer) に組み込めるようにし、同時に (Apple) は (Windows) の (マルチメディア) (技術) の (サポート) を (強化) することになろう。	0.0	2.2	72
今までは (Windows) 用の (QuickTime) は (再生) しかできなかった。	2.2	1.1	74

【図９】


文選択処理の詳細フローチャート
文選択処理開始
注目概念リストを作成
S20
選択結果リストをクリア
S21
S22
必須出力句リストが空か?
No
S23
先頭の句を必須出力句リストから除き、選択結果リストに追加
S24
選択結果リストに追加した句中の事象を既知事象リストに追加
Yes
S25
全ての述語句(必須出力句以外)を選択候補リストに追加
選択候補リスト中の全ての句中の注目情報量を求める
S26
注目情報量が0の句を選択候補リストから除く
S27
選択候補リスト中の全ての句の新情報量を求める
S28
新情報量が0の句を選択候補リストから除く
S29
No
選択候補句リストが空でない?
S30
Yes
注目情報量最大の句の中で新情報量最大の句を選び、出力句とする
S31
出力句を選択候補リストから除き、選択結果リストに追加
S32
出力句中の注目概念を注目概念リストから除き、注目情報量を再計算する
S33
注目情報量が0の句を選択候補リストから除く
S34
出力句中の事象を既知事象リストに追加し、新情報量を再計算する
S35
新情報量が0の句を選択候補リストから除く
S36
文選択処理終了

【図１１】

第１と第２の新情報量を区別する場合
の新情報量の比較処理フローチャート



【図１９】

文選択方式の他の実施例としての
ダイジェスト情報抽出のアルゴリズムを示す図

未出現キーワード列　←見出しキーワード。
抽出結果　←空。
ｗｈｉｌｅ未出現キーワード列が空でない。
　ｄｏ（重要文選択）
　　未出現キーワード列に基づき文の重要
　　度を評価。
　　重要度最大の文（複数可）を選択。
　　（１つも選べなければ処理終了。）
　　選択された文を抽出結果に加える。
　　選択された文中の見出しキーワードを
　　未出現キーワード列から除く。
　ｏｄ

【図２０】

図１９のアルゴリズムと図９のフローチャートとにおける用語の対応を示す図

図１９	図９
未出現キーワード列	注目概念リスト
抽出結果	選択結果リスト
重要度	注目情報量と新情報量
(1) 見出しキーワードと一致する名詞の異なり数	(第一注目情報量)
(2) 見出しキーワードと一致する名詞の延べ数	(第二注目情報量)
(3) 見出しキーワードと一致しない名詞の延べ数	(新情報量)

【図１２】

特許抄録の要約結果の例を示す図

公開番号	特開平 06-259424
発明の名称	文書表示装置及び文書要約装置並びにデイジタル複写装置
発明者	亀田　雅之 (リコー)
抄録抜粋	…本発明による【文書】【表示】【装置】は、見出し部分と本文とから成る【文書】に対する【表示】機能を有するもので、解析手段１は見出し部分を解析し、認識手段２は、前記解析手段１により解析された見出し部分中の単語を本文中から認識する。…
特徴語	装置, 表示, 文書
公開番号	特開平 07-036896
発明の名称	文書を要約する方法および装置
発明者	エム　マーガレット　ウイズゴット／ダグラス　アール　カッテイング (ゼロックス)
抄録抜粋	…本発明の【方法】は、【文書】の領域を選択することによつて【要約】を自動的に作成する。…
特徴語	文書, 方法, 要約
公開番号	特開平 08-297677
発明の名称	主題の要約を生成する自動的な方法
発明者	フランシーヌ　アール　チエン (ゼロックス)
抄録抜粋	【課題】機械で読み取り可能なドキユメントの【主題】の【要約】を自動的に【生成】する【方法】を提供する。…
特徴語	主題, 生成, 方法, 要約
公開番号	特開平 06-215049
発明の名称	文書要約装置
発明者	乾　隆夫／芥子　育雄／石鉢　康一郎 (シャープ)
抄録抜粋	…文書処理部４は、各文章ベクトル間距離を参照して、【文書】に最も近い段落と【文書】に近い複数文との２種類の要旨及び【文書】に最も近い各段落毎の文と各段落に最も近い文との２種類の【要約】を生成する。…
特徴語	文書, 要約
公開番号	特開平 07-182373
発明の名称	文書情報検索装置及び文書検索結果表示方法
発明者	住田　一男／三池　誠司／小野　顕司／竹林　洋一／武田　公人／伊藤　悦雄 (東芝)
抄録抜粋	【検索】結果を原文で【表示】するのではなく、利用者の所望する視点での【要約】文を提示することにより、利用者が【検索】した【文書】の内容をたやすく理解し、要不要をすばやく判定することを可能にする【文書】【情報】【検索】【装置】を提供する。…
特徴語	検索, 情報, 装置, 表示, 文書, 要約
公開番号	特開平 07-044566
発明の名称	抄録作成装置
発明者	小野　顕司／住田　一男／三池　誠司 (東芝)
抄録抜粋	（ｊ）【抄録】文中の照応表現の照応先或いは省略箇所表示を【抄録】文中に含ませることにより、【抄録】文の読解性、自然性の向上を図る。書式解析部１は電子化された入力【文章】を解析し、文の切れ目、段落の変わり目、章や節の構造等を解析する。…
特徴語	抄録, 文章

【図１５】

図14において文72が選択された後の情報量を示す図

文番号	注目情報量	新情報量	文
1	1.1	9.9	〔G. Amelio〕は〔社内〕の〔機構〕を〔改革〕し、〔Macintosh〕の〔機種〕を半分に減らして〔開発〕費を減らし、約3,000人の〔社員〕を〔レイオフ〕しつつ、【AppleComputre】を建て直している。
3	1.1	3.3	これは【Apple】【再建】に最も〔重要〕な〔地位〕である。
11	1.1	8.9	しかし〔Hancock〕は〔ソフトウェア〕をよく知っており、〔Apple〕【再建】の〔成否〕は〔開発〕が遅れ続けている〔Copland〕にかかっているので、〔Hancock〕は〔妥当〕な〔人事〕と見られている。
21	1.1	5.5	〔Amelio〕は〔Apple〕【再建】に〔妥当〕な〔手〕を打っていると見られるが、結果が〔目〕に見えてくるまでは少なく共1年はかかるであろう。
41	0.0	6.6	〔Apple〕は〔Macintosh〕〔互換〕機〔戦略〕を【強化】しつつあるが、そのための〔製品〕を夏から〔出荷〕する。
72	・・・		(選択ずみ)
74	0.0	2.2	今までは【Windows】用の〔QuickTime〕は〔再生〕しかできなかった。

【図１６】

図13(b)の結果を得るための依存関係の設定を説明する図

文番号	注目情報量	新情報量	文
1	1.1	9.9	〔G. Amelio〕は〔社内〕の〔機構〕を〔改革〕し、〔Macintosh〕の〔機種〕を半分に減らして〔開発〕費を減らし、約3,000人の〔社員〕を〔レイオフ〕しつつ、【AppleComputre】を建て直している。
2	0.0	11.11	〔Amelio〕は〔Apple〕の〔重要〕な〔地位〕に外部から〔人〕を入れているが、〔研究開発〕の〔最高〕の〔担当者〕〔chief technology officer〕として53才の〔EllenHancock〕を〔任命〕した。
3	1.1	3.3	これは【Apple】【再建】に最も〔重要〕な〔地位〕である。
11	1.1	8.9	しかし〔Hancock〕は〔ソフトウェア〕をよく知っており、〔Apple〕【再建】の〔成否〕は〔開発〕が遅れ続けている〔Copland〕にかかっているので、〔Hancock〕は〔妥当〕な〔人事〕と見られている。
21	1.1	5.5	〔Amelio〕は〔Apple〕【再建】に〔妥当〕な〔手〕を打っていると見られるが、結果が〔目〕に見えてくるまでは少なく共1年はかかるであろう。

(a) 依存関係にある部分

文1 ← 依存《〔Amelio〕は》← 文2 ← 依存《これは》← 文3

(b) 文3に関する依存関係

文1 ← 依存《〔Amelio〕は》← 文2 ← 依存《〔Hancock〕は》← 文11

(c) 文11に関する依存関係

文1 ← 依存《〔Amelio〕は》← 文21

(d) 文21に関する依存関係

【図１７】

図16の依存関係を考慮した情報量の計算結果を示す図

文番号	注目情報量	新情報量	文
1	1.1	9.0	(G. Amelio)は(出身)の(創業)を(過去)し、(Macintosh)の(開発)を中心に進らして(開発)費を膨らし、約5,000人の(社員)を(レイオフ)しつつ、【AppleComputer】を建て直している。
3(=211)	2.4	15.21	これは【Apple】【再建】に進も(重要)な(施策)である。
11(=211)	2.4	24.27	しかし(Hancock)は(ソフトウェア)をよく知っており、(Apple)【再建】の(戦略)は(開発)が遅れている(Copland)にかかっているので、(Hancock)は(注目)な(人事)と見られている。
21(=1)	2.3	12.13	(Amelio)は(Apple)【再建】に(重当)な(手)を打っていると見られるが、結果が(目)に見えてくるまでは少なく(※1年以内)かもであろう。
41	1.1	6.6	(Apple)は(Macintosh)(互換)機(開発)を(強化)しつつあるが、そのための(資金)を更から(出府)する。
72	2.2	8.8	(交換)がなされば(Macintosh)は【QuickTime】を(Internet Explorer)に組み込めるようにし、同時に(Apple)は【Windows】の(マルチメディア)(技術)の(サポート)を(強化)することになろう。
74	1.1	2.2	今までは【Windows】用の(QuickTime)は(再生)しかできなかった。

【図１８】

図17において文11が選択された後の情報量を示す図

文番号	注目情報量	新情報量	文
1	(選択ずみ)		
8	0.0	3.3	これは【Apple】【再建】に進も(重要)な(施策)である。
11(=211)	(選択ずみ)		
21	0.0	5.5	(Amelio)は(Apple)【再建】に(重当)な(手)を打っていると見られるが、結果が(目)に見えてくるまでは少なく(※1年以内)かもであろう。
41	1.1	6.6	(Apple)は(Macintosh)(互換)機(開発)を(強化)しつつあるが、そのための(資金)を更から(出府)する。
72	2.2	8.8	(交換)がなされば(Macintosh)は【QuickTime】を(Internet Explorer)に組み込めるようにし、同時に(Apple)は【Windows】の(マルチメディア)(技術)の(サポート)を(強化)することになろう。
74	1.1	2.2	今までは【Windows】用の(QuickTime)は(再生)しかできなかった。